numero 186
febbraio 1984
anno XVII
volume XXXII

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

La gravità quantistica

In una teoria quantomeccanica della gravitazione la stessa geometria dello spazio e del tempo sarebbe soggetta a continue fluttuazioni e perfino la distinzione tra passato e futuro potrebbe divenire incerta

di Bryce S. DeWitt

Tra le forze della natura la gravità pare abbia uno stato particolare. Altre forze, come l'elettromagnetismo, agiscono *nello* spazio-tempo, che ha semplicemente la funzione di riferimento per gli eventi fisici. La gravità è completamente diversa: non è una forza applicata su un fondo passivo di spazio e di tempo, ma costituisce una distorsione dello spazio-tempo stesso. Un campo gravitazionale è una «curvatura» dello spazio-tempo. È questa la concezione della gravità che Einstein raggiunse in quella che descrisse come la più pesante fatica della sua vita.

La distinzione qualitativa tra la gravità e le altre forze diventa molto più chiara quando si tenta di formulare una teoria della gravitazione che concordi con i precetti della meccanica quantistica. Il mondo quantistico non è mai in quiete. Per esempio, nella teoria quantistica dei campi elettromagnetici, il valore del campo elettromagnetico fluttua continuamente. In un universo dominato dalla gravità quantistica sarebbero soggette a fluttuazioni la curvatura dello spazio-tempo e perfino la sua stessa struttura. È probabile in realtà che la sequenza degli eventi nel mondo e il significato di passato e di futuro siano suscettibili di modificazioni.

Qualcuno potrebbe pensare che, se esistessero fenomeni del genere, sicuramente dovrebbero già essere stati rilevati. Accade, invece, che tutti gli effetti di natura quantomeccanica della gravitazione siano confinati in una scala straordinariamente piccola, sulla quale, nel 1899, Max Planck richiamò per primo l'attenzione. In quell'anno, Planck introdusse la sua famosa costante, chiamata quanto d'azione e indicata con $\hbar$. Egli stava cercando di dare un significato allo spettro della radiazione di corpo nero, la luce che sfugge da una piccola apertura praticata in una cavità molto calda. Come fatto curioso, notò che la sua costante, combinata con la velocità della luce e con la costante di gravitazione di Newton, dà origine a un sistema assoluto di unità di misura. Tali unità forniscono la scala della gravità quantistica.

Le unità di Planck sono completamente estranee alla fisica di ogni giorno. Per esempio, l'unità di lunghezza è di $1,61 \times 10^{-33}$ centimetri, ovvero inferiore di 21 ordini di grandezza al diametro di un nucleo atomico. Essa sta alle dimensioni nucleari grosso modo nello stesso rapporto in cui stanno le dimensioni dell'uomo a quelle della nostra galassia. Ancora più curiosa è l'unità di tempo di Planck: $5,36 \times 10^{-44}$ secondi. Per verificare sperimentalmente queste scale di distanza e di tempo impiegando strumenti costruiti con l'attuale tecnologia sarebbe necessario un acceleratore di particelle delle dimensioni della Galassia!

Dal momento che la via sperimentale non ci può aiutare, la gravità quantistica è insolitamente speculativa. Ciononostante, essa è di spirito fondamentalmente conservatore: prende la teoria attualmente consolidata e si limita a spingerla fino alle sue estreme conseguenze logiche. Nei suoi aspetti essenziali ha per obiettivo quello di fondere tre teorie: la relatività ristretta, la teoria einsteiniana della gravitazione e la meccanica quantistica, e nient'altro. Una tale sintesi non è stata ancora completamente realizzata, ma nel tentativo di raggiungerla si è già potuto apprendere molto.

Lo sviluppo di una valida teoria della gravità quantistica offre, inoltre, la sola strada che si conosca verso la conoscenza dell'origine del big bang e del destino finale dei buchi neri, eventi che si possono considerare caratteristici dell'inizio e della fine dell'universo.

Delle tre teorie che convergono nella gravità quantistica, la relatività ristretta è venuta storicamente per prima. È la teoria che unisce spazio e tempo attraverso il postulato (poi confermato sperimentalmente) che la velocità della luce è la stessa per tutti gli osservatori che si muovono nel vuoto, sottratti a forze esterne. Le conseguenze di questo postulato, introdotto nel 1905 da Einstein, si possono descrivere con l'aiuto di un diagramma spazio-tempo, un grafico che riporta curve che rappresentano le posizioni di oggetti nello spazio in funzione del tempo. Le curve sono chiamate «linee universali».

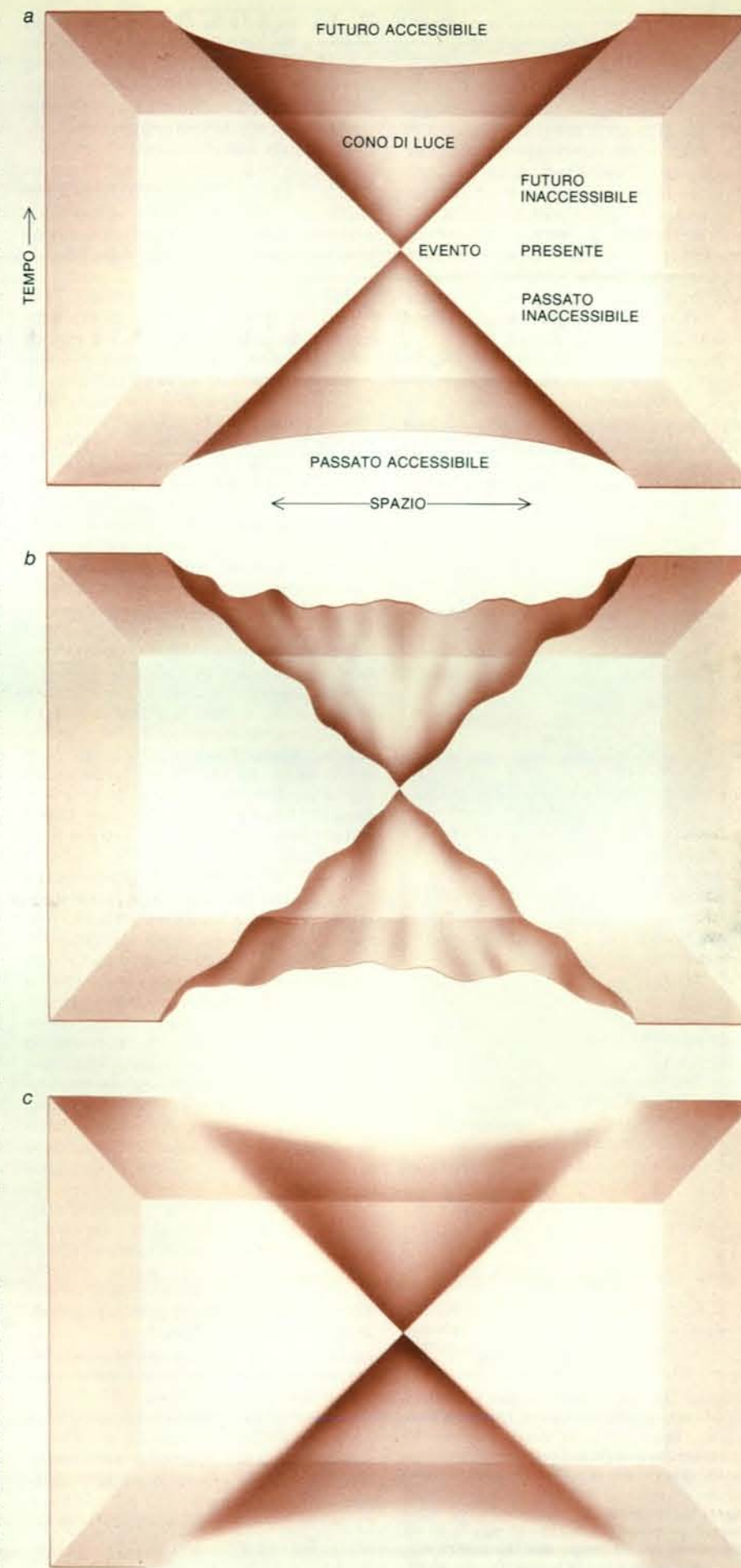
Per amore di semplicità ignorerò due delle dimensioni spaziali. Si può allora tracciare una linea universale su un grafico bidimensionale nel quale si misurano orizzontalmente le distanze spaziali e verticalmente gli intervalli di tempo. Una retta verticale è la linea universale di un oggetto in quiete nel sistema di riferimento scelto per la misurazione. Una retta inclinata è la linea universale di un oggetto in moto a velocità costante nel sistema di riferimento scelto. Una linea universale curva rappresenta, infine, un oggetto sottoposto ad accelerazione.

Un punto del diagramma spazio-tempo definisce sia una posizione dello spazio sia un istante di tempo ed è chiamato evento. La distanza spaziale tra due eventi dipende dal sistema di riferimento prescelto e lo stesso vale per l'intervallo di tempo. Il concetto di simultaneità dipende dal sistema di riferimento: due eventi collegati da una linea orizzontale in un dato sistema di riferimento sono simultanei in tale sistema, ma non in altri.

Per stabilire una relazione tra sistemi di riferimento in moto relativo, si deve introdurre un'unità comune per la misura dello spazio e del tempo. La velocità della luce giunge da fattore di conversione, collegando una data distanza al tempo necessario perché la luce la percorra. Adotterò il metro come unità sia dello spazio sia del tempo. Un metro di tempo è pari a circa 3,33 nanosecondi (miliardesimi di secondo).

Misurando lo spazio e il tempo nelle stesse unità, la linea universale di un fotone (un quanto di luce) è inclinata a 45 gradi. La linea universale di qualsiasi oggetto materiale ha, invece, un'inclinazione rispetto alla verticale sempre minore di 45 gradi, il che è un altro modo di dire che la sua velocità è sempre inferiore a quella della luce. Se la linea universale di un oggetto o di un segnale qualsiasi fosse inclinata a più di 45 gradi dalla verticale, a certi osservatori l'oggetto o il segnale apparirebbe muoversi a ritroso nel tempo. Mettendo a punto un relè per segnali più veloci della luce, un uomo potrebbe trasmettere informazioni nel suo passato, violando in tal modo il principio di causalità. Tali segnali sono però vietati dalle caratteristiche della relatività ristretta.

Si considerino due eventi sulla linea universale di un osservatore non sottoposto ad accelerazione. Si supponga che gli eventi, in un particolare sistema di riferimento, siano distanti quattro metri nello spazio e cinque metri nel tempo. In tale sistema l'osservatore si sta quindi muovendo ai quattro quinti della velocità della luce. In un altro sistema la sua velocità sarebbe differente e la stessa cosa accadrebbe per gli intervalli di spazio e di tempo associati. C'è però una grandezza che si manterrebbe inalterata in tutti i sistemi di riferimento. Questa grandezza invariante è detta «tempo proprio» tra i due eventi ed è

Il cono di luce, che definisce le regioni dell'universo accessibili da un dato punto dello spazio e da un dato istante di tempo, diventerebbe un concetto male espresso in una teoria della gravità quantistica. Il cono (*a*) è una superficie nello spazio-tempo tetradimensionale, ma viene qui rappresentato eliminando una dimensione spaziale. Se la gravitazione è quantizzata, la forma del cono può fluttuare fortemente su brevi distanze (*b*). In realtà le fluttuazioni non si possono percepire direttamente; il cono di luce si comporta come se fosse vago. Alla domanda se due punti dello spazio-tempo possano comunicare l'uno con l'altro (mediante segnali in moto a velocità inferiore a quella delle luce) si può quindi dare solo una risposta probabilistica (*c*).

l'intervallo di tempo misurato da un orologio che l'osservatore porta con sé.

Nel sistema di riferimento prescelto la linea universale tra i due eventi è l'ipotenusa di un triangolo rettangolo avente una base di quattro metri e un'altezza di cinque. Il tempo proprio corrisponde alla «lunghezza» di quest'ipotenusa, ma viene calcolato in modo insolito: mediante un «teorema pseudopitagorico». Come nel caso del normale teorema di Pitagora, si cominciano a calcolare i quadrati dei lati del triangolo. Nella relatività ristretta, però, il quadrato dell'ipotenusa è uguale alla differenza tra i quadrati dei cateti anziché alla loro somma.

Nel presente esempio il tempo proprio è di tre metri e resta di tre metri nel sistema di riferimento di qualsiasi osservatore non sottoposto ad accelerazione. Questa invarianza del tempo proprio è ciò che unisce spazio e tempo in un'unica entità.

La geometria dello spazio-tempo, essendo basata su un teorema pseudopitagorico, non è quella euclidea, ma per molti aspetti è analoga a essa. Nella geometria euclidea, tra tutte le linee che collegano due punti una retta si può definire come linea di lunghezza estrema. Lo stesso vale per la geometria dello spazio-tempo. Nella geometria euclidea, però, l'estremo è sempre un minimo, mentre nello spazio-tempo è un massimo quando i due punti si possono collegare mediante una linea universale che richiede un viaggio a velocità non superiore a quella della luce.

La linea universale definisce una traiettoria attraverso lo spazio e il tempo. Qui sono indicate due linee universali in una versione del paradosso dei gemelli di Einstein. La linea universale «curva» del gemello che subisce accelerazione nel punto di inversione del suo viaggio appare la più lunga, ma tale gemello registra il «tempo proprio» più breve. In effetti, in un diagramma spazio-tempo una linea retta rappresenta l'intervallo più lungo tra due punti.

Nel 1854 il matematico tedesco G. F. B. Riemann trovò una generalizzazione della geometria euclidea per gli spazi curvi. Due spazi curvi bidimensionali sono stati studiati fin dall'antichità: essi sono chiamati superfici curve e sono solitamente visti nella prospettiva dello spazio euclideo tridimensionale ordinario. Riemann dimostrò che uno spazio curvo può avere un numero di dimensioni qualsiasi e che può essere studiato intrinsecamente. Non è necessario che lo si immagini immerso in uno spazio euclideo con un maggior numero di dimensioni.

Riemann sottolineò, inoltre, che lo spazio fisico in cui viviamo può essere curvo. Secondo Riemann, la questione si potrebbe decidere soltanto con un esperimento. Come si potrebbe eseguire un siffatto esperimento, almeno in linea di principio? Si dice che lo spazio euclideo è piatto. Uno spazio piatto ha la proprietà che si possono tracciare rette parallele in modo da creare una griglia rettangolare uniforme. Che cosa accadrebbe se si tentasse di disegnare una griglia come questa sulla superficie della Terra, supponendo che la Terra sia piatta?

Si può osservare il risultato da un aereo in un giorno limpido, al di sopra delle regioni coltivate delle Great Plains americane. Il territorio viene suddiviso da strade che corrono da est a ovest e da nord a sud in sezioni di un miglio quadrato. Le strade che corrono da est a ovest si estendono spesso ininterrottamente per molte miglia, ma non quelle che corrono da nord a sud. Percorrendo una strada verso il nord vi sono ogni poche miglia brusche svolte verso est o verso ovest che sono dovute alla curvatura della Terra. Se si eliminassero, le strade confluirebbero fino a creare sezioni di meno di un miglio quadrato.

Nel caso di uno spazio tridimensionale, si può pensare di costruire in esso un'impalcatura gigante fatta di tubolari rettilinei di uguale lunghezza congiunti in modo da formare angoli esattamente di 90 e di 180 gradi. Se lo spazio è piatto, la costruzione dell'impalcatura procederebbe senza difficoltà; se è curvo, prima o poi sarà inevitabile dover accorciare o tirare i tubolari per farli combaciare.

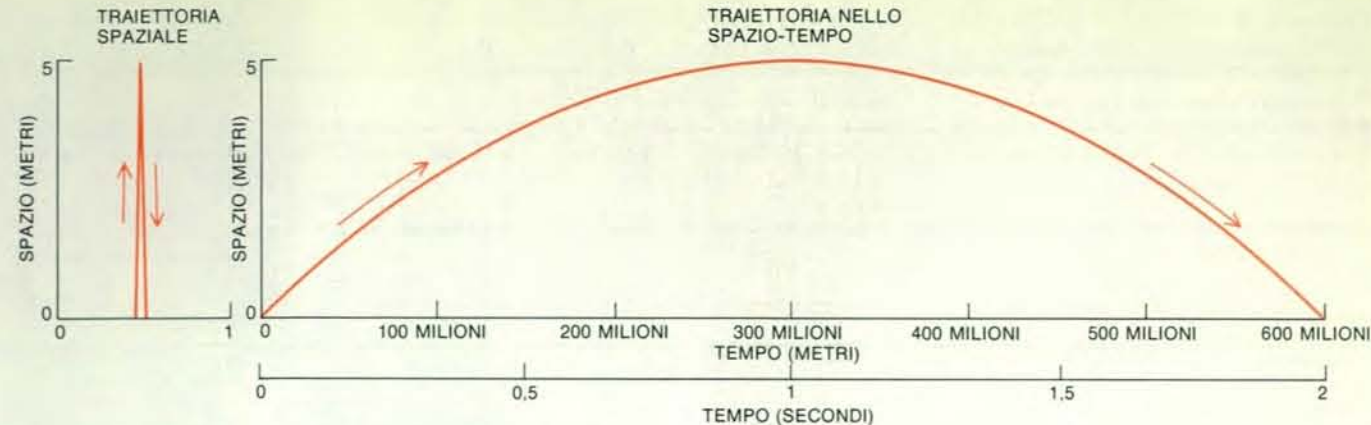
La stessa generalizzazione introdotta da Riemann nella geometria euclidea si può applicare alla geometria della relatività ristretta. La generalizzazione fu operata da Einstein tra il 1912 e il 1915 con l'aiuto del matematico Marcel H. Grossmann. Il risultato è una teoria dello spazio-tempo curvo. In mano a Einstein si trasformò in una teoria della gravitazione. Nella relatività ristretta i campi gravitazionali si considerano assenti e lo spazio-tempo si assume piatto. In uno spazio-tempo curvo è presente un campo gravitazionale: in realtà, «curvatura» e «campo gravitazionale» sono sinonimi.

Dal momento che la teoria del campo gravitazionale di Einstein è una generalizzazione della relatività ristretta, egli la chiamò relatività generale. Il nome è improprio. La relatività generale è in realtà meno relativistica della relatività ristretta. La completa mancanza di peculiarità dello spazio-tempo piatto, la sua omogeneità e isotropia sono ciò che garantisce che le posizioni e le velocità siano strettamente correlate. Non appena lo spazio-tempo si arricchisce di «protuberanze», cioè regioni locali di curvatura, diventa assoluto perché si possono precisare posizione e velocità rispetto alle protuberanze. Lo spazio-tempo, invece di essere semplicemente un'arena priva di caratteristiche peculiari per la fisica, è a sua volta dotato di proprietà fisiche.

Nella teoria di Einstein la curvatura è prodotta dalla materia. La relazione tra la quantità di materia e il grado di curvatura è semplice in linea di principio, ma complicata da calcolarsi. Sono necessarie venti funzioni delle coordinate di un punto dello spazio-tempo per descrivere la curvatura in quel punto. Dieci di tali funzioni corrispondono a una parte della curvatura che si propaga liberamente sotto forma di onde gravitazionali, o «oscillazioni di curvatura». Le altre 10 funzioni sono determinate dalla distribuzione della massa, della quantità di moto, del momento angolare e dalle tensioni interne della materia, nonché dalla costante di gravitazione di Newton, G.

Con riferimento alle densità di massa incontrate sulla Terra, G è una costante piccolissima. È necessaria una massa enorme per curvare apprezzabilmente lo spazio-tempo. La grandezza inversa $1/G$ si può considerare come una misura della «rigidità» dello spazio-tempo. In base all'esperienza quotidiana, lo spazio-tempo è molto rigido. L'intera massa della Terra induce una curvatura dello spazio-tempo che è pari a solo un miliardesimo circa della curvatura della superficie terrestre.

Nella teoria di Einstein un corpo in caduta libera o un corpo liberamente orbitante seguono una linea universale geodetica. Una geodetica che collega due punti dello spazio-tempo è una linea universale di lunghezza estrema tra essi: è una generalizzazione del concetto di linea retta. Se si immagina uno spazio-tempo curvo immerso in uno spazio piatto di maggior numero di dimensioni, una geodetica appare come una linea curva.

L'effetto della curvatura su un corpo in movimento è stato spesso illustrato da un modello nel quale una sfera rotola su un foglio di gomma deformato. Il modello non è esatto in quanto può rappresentare soltanto la curvatura spaziale. Nella vita reale siamo costretti a restare nell'universo a quattro dimensioni dello spazio e del

La curvatura dello spazio-tempo in presenza di una massa costituisce un campo gravitazionale. Una palla, lanciata in aria a un'altezza di cinque metri, resta sollevata per due secondi. Il suo moto di salita e di discesa rivela la curvatura dello spazio-tempo nei pressi della superficie terrestre. La curvatura della traiettoria è immediatamente visibile, ma è in realtà piccolissima quando si misurano lo spazio e il tempo nelle stesse unità. Per esempio, i secondi si possono trasformare in metri moltiplicandoli per la velocità della luce, pari a 300 milioni di metri al secondo. In tal caso, la traiettoria diventa un arco estremamente piatto lungo 600 milioni di metri e alto cinque metri (*a destra*).

tempo. Inoltre, non possiamo evitare di muoverci in tale universo, perché siamo proiettati in avanti nel tempo. Il tempo è l'elemento chiave. Risulta che, benché in un campo gravitazionale lo spazio sia curvo, è molto più importante la curvatura del tempo. Ciò è dovuto all'elevato valore della velocità della luce, che collega la scala dello spazio a quella del tempo.

Vicino alla Terra la curvatura dello spazio è talmente lieve da non potersi rilevare con misurazioni statiche. Tuttavia la nostra precipitosa corsa nel tempo è così veloce che nelle situazioni dinamiche la curvatura diventa notevole, allo stesso modo in cui una lieve gobba in un'autostrada, pur passando inosservata a un pedone, può diventare pericolosa per un'automobile veloce. Lo spazio attorno alla Terra appare piatto, con un alto grado di precisione, ma possiamo vedere la curvatura dello spazio-tempo semplicemente lanciando in aria una palla. Se la palla rimane in aria per due secondi, descrive un arco con un'altezza di cinque metri. La luce percorre 600 000 chilometri in due secondi. Se immaginiamo l'arco alto cinque metri stirato orizzontalmente fino a una lunghezza di 600 000 chilometri, la curvatura dell'arco è la curvatura dello spazio-tempo.

L'introduzione da parte di Riemann del concetto di spazi curvi diede l'avvio a un'altra fruttuosa branca della matematica: la topologia. Si sapeva che esistono superfici bidimensionali prive di contorni in una varietà infinita di tipi che non possono essere deformati l'uno nell'altro con continuità; ne sono due semplici esempi una sfera e un toro. Riemann fece notare che la stessa cosa vale per spazi curvi con un maggior numero di dimensioni e fece i primi passi per una loro classificazione.

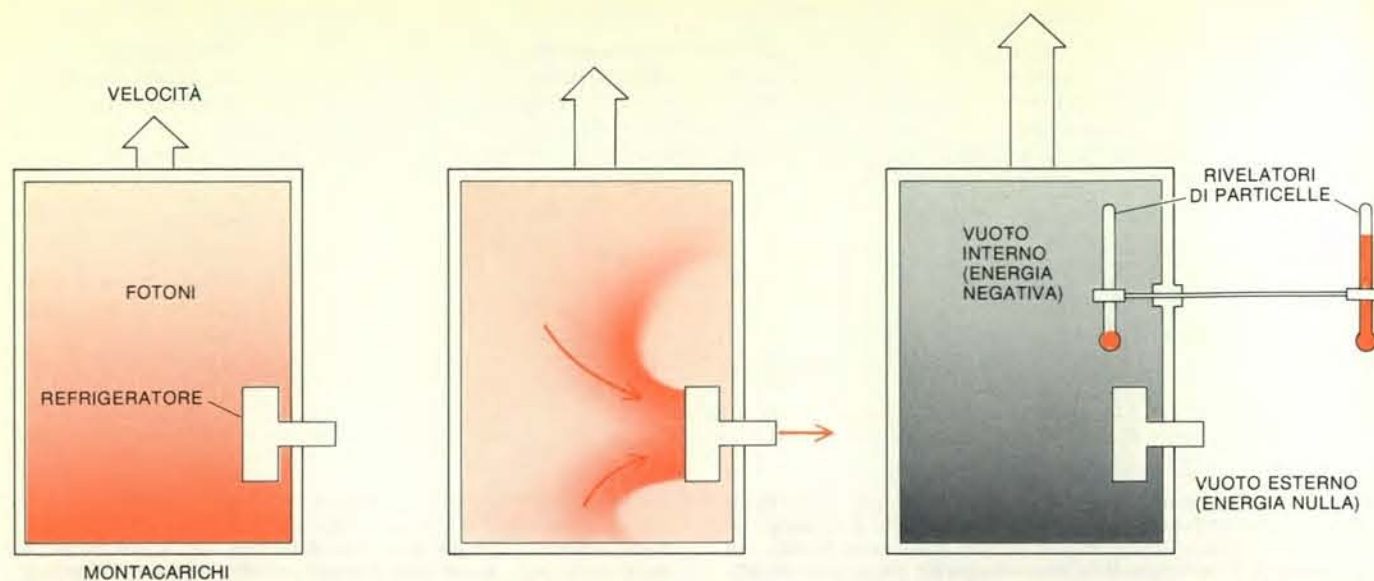
Esiste un numero infinito di tipi topologici anche dello spazio-tempo curvo (o, più esattamente, dei modelli di spazio-tempo curvo). Alcuni modelli si possono rifiutare come candidati per una descrizione dell'universo reale perché portano a paradossi di causalità o perché in essi non possono essere rispettate le leggi fisiche note. Tuttavia resta ancora un numero di possibilità enorme.

Un modello dell'universo degno di nota venne proposto dal matematico russo Alexander A. Friedmann nel 1922. Nella relatività ristretta lo spazio-tempo viene visto non solo come piatto, ma anche di estensione infinita sia nello spazio sia nel tempo. Nel modello di Friedmann qualsiasi sezione trasversale spaziale a tre dimensioni dello spazio-tempo ha un volume finito e ha la topologia di una trisfera, uno spazio che può essere immerso in uno spazio euclideo quadridimensionale in modo tale che tutti i suoi punti siano equidistanti da un punto dato. Il modello è stato il preferito dai cosmologi dal momento in cui Edwin P. Hubble, negli anni venti, scoprì l'espansione dell'universo. Se si combina il modello di Friedmann con la teoria della gravitazione di Einstein, esso prevede un big bang in un istante iniziale di compressione infinita, seguito da un'espansione che dura miliardi di anni per mutua attrazione gravitazionale di tutta la materia dell'universo.

Uno spazio-tempo di Friedmann ha la proprietà che qualsiasi curva chiusa disegnata in esso può essere ridotta in modo continuo a un punto. Uno spazio-tempo siffatto è detto «semplicemente connesso». L'universo reale non può avere una tale proprietà. Pare che il modello di Friedmann descriva molto bene la regione di spazio entro alcuni miliardi di anni luce nella nostra galassia, ma non possiamo vedere l'intero universo.

Un semplice esempio di universo molteplicemente connesso è quello di una struttura ripetuta all'infinito, come il motivo di una carta da parati, in una data direzione spaziale. Ogni galassia di un siffatto universo è un membro di una serie infinita di galassie identiche poste a una distanza fissa (e necessariamente enorme) l'una dall'altra. Se i membri di una serie sono veramente identici, è discutibile se si debbano considerare distinti. È più conveniente considerare ogni serie come rappresentante una sola galassia. Un viaggio da un membro della serie a quello successivo riporta, quindi, un viaggiatore al punto di partenza e una linea che descrive tale viaggio è una curva chiusa che non può essere ridotta a un punto. Essa assomiglia a una curva chiusa sulla superficie di un cilindro che gira attorno al cilindro una sola volta. L'universo che si ripete è detto universo cilindrico.

Un altro esempio di una struttura molteplicemente connessa, su una scala molto più piccola, è il *wormhole* (alla lettera cunicolo o galleria di tarlo), introdotto nel 1957 da John Archibald Wheeler, ora all'Università del Texas ad Austin. Si può costruire un «cunicolo» bidimensionale praticando due aperture circolari in una superficie bidimensionale e congiungendone accuratamente i bordi (*si veda l'illustrazione a pagina 13*). Il procedimento nelle tre dimensioni è lo stesso, ma è più difficile visualizzarlo.

Dal momento che le due aperture possono trovarsi a una grande distanza nello spazio originario, anche se avvicinate dal passaggio che le collega, il cunicolo è diventato un dispositivo comune nella fantascienza per spostarsi da un punto a un altro molto più velocemente di quanto possa fare la luce: basta praticare due aperture nello spazio, collegarle e strisciare lungo il passaggio. Sfortunatamente, anche se si potesse costruire un perforatore (il che è molto dubbio), lo schema non funzionerebbe. Se la geometria dello spazio-tempo è regolata dalle equazioni di Einstein, il cunicolo è un oggetto dinamico. Ne consegue che le due aperture che esso collega sono necessariamente buchi neri e qualsiasi cosa entri in esse non ne può più uscire. Ciò che avviene è che il passaggio «si restringe» e che ogni cosa al suo interno viene compressa a una densità infinita prima di poter raggiungere l'altro capo.

Un montacarichi è l'apparecchiatura adatta per un esperimento ideale che si basa sulla natura del vuoto nella meccanica quantistica e sull'effetto che l'accelerazione o la gravitazione hanno sul vuoto. Si suppone che il montacarichi sia vuoto e sigillato, in modo che inizialmente esiste un vuoto perfetto sia all'interno sia all'esterno del montacarichi. Appena inizia l'accelerazione, però, viene emessa un'onda elettromagnetica dal pavimento e il montacarichi si riempie di un tenue gas di fotoni, o quanti di radiazione elettromagnetica (*a sinistra*). Un refrigeratore alimentato da una fonte di energia esterna estrae fotoni (*al centro*). Una volta eliminati tutti i fotoni, i rivelatori di fotoni misurano l'energia del vuoto sia all'interno sia all'esterno (*a destra*). Poiché lo strumento all'esterno sta accelerando nel vuoto, esso è sensibile alle fluttuazioni quantomeccaniche dei campi che permeano lo spazio anche in assenza di particelle. Il rivelatore all'interno è in quiete rispetto al montacarichi e non sente le fluttuazioni. Ne consegue che i vuoti all'interno e all'esterno del montacarichi non sono equivalenti. Se si attribuisce energia nulla al vuoto «standard» all'esterno del montacarichi, il vuoto all'interno deve avere energia negativa. Per poter riportare l'energia a zero, si dovrebbero ripristinare i fotoni rimossi dal refrigeratore. Anche un campo gravitazionale può creare un vuoto con energia negativa.

La meccanica quantistica, la terza componente della gravità quantistica, è stata ideata nel 1925 da Werner Heisenberg e da Erwin Schrödinger, ma la loro formulazione iniziale non teneva conto della teoria della relatività. Il suo successo fu ciononondimeno immediato e brillante, perché attendevano di essere spiegate moltissime osservazioni sperimentali nelle quali dominano gli effetti quantistici, mentre la relatività ha un ruolo di minore importanza o trascurabile. Si sapeva però che in alcuni atomi gli elettroni raggiungono velocità pari a una notevole frazione della velocità della luce e, quindi, la ricerca di una teoria quanto-relativistica non venne rinviata a lungo.

Alla metà degli anni trenta era già chiaro che, quando si combina la teoria quantistica con la relatività, si possono dedurre numerosi fatti del tutto nuovi, fra i quali due di fondamentale importanza. In primo luogo, ogni particella è associata a un tipo di campo e ogni campo è associato a una classe di particelle indistinguibili. Non fu più possibile considerare il campo elettromagnetico e quello gravitazionale come i soli campi fondamentali della natura. In secondo luogo, esistono due tipi di particelle classificate secondo il loro momento angolare di spin (quantizzato). Quelle con spin $1/2\hbar$, $3/2\hbar$ e così via seguono il principio di esclusione (non possono coesistere due particelle nello stesso stato quantico); quelle con spin 0, $\hbar$, $2\hbar$ e così via sono gregarie.

Queste sorprendenti conseguenze derivanti dall'unione della relatività ristretta alla meccanica quantistica sono state ripetutamente confermate nell'ultimo mezzo secolo. La relatività e la teoria dei quanti insieme conducono a una teoria superiore alla somma delle due parti. L'effetto sinergico è ancora più pronunciato allorché si inserisce la gravità.

Nella fisica classica uno spazio-tempo piatto e vuoto («il vuoto» per eccellenza) è privo di strutture, mentre nella fisica quantistica il nome di «vuoto» è dato a un'entità più complessa dotata di una ricca struttura. La sua struttura deriva dall'esistenza nel vuoto di campi liberi che non si annullano mai, campi, cioè, lontani dalle loro sorgenti.

Un campo elettromagnetico libero è matematicamente equivalente a un insieme infinito di oscillatori armonici, che si possono rappresentare come molle alle quali sono fissate delle masse. Nel vuoto ciascun oscillatore si trova nel suo stato fondamentale, o stato di minima energia. Quando un oscillatore classico (non quantomeccanico) si trova nel suo stato fondamentale, è immobile in un punto ben definito. Ciò non è vero per un oscillatore quantistico. Se un oscillatore quantistico fosse in un punto ben definito, la sua posizione sarebbe nota con precisione infinita; per il principio di indeterminazione allora dovrebbe avere quantità di moto ed energia infinite, il che è impossibile. Nello stato fondamentale di un oscillatore quantistico non sono esattamente definite né la posizione né la quantità di moto. Entrambe sono soggette a fluttuazioni casuali. Nel vuoto quantistico è il campo elettromagnetico (e qualsiasi altro campo) a fluttuare.

Benché casuali, le fluttuazioni del campo nel vuoto quantistico sono di una specie particolare. Soddisfano il principio di relatività nel senso che «paiono» le stesse a tutti gli osservatori non accelerati, qualunque sia la loro velocità. Si può dimostrare che questa proprietà implica che il campo sia nullo in media e che le fluttuazioni aumentino di ampiezza alle lunghezze d'onda minori. Il risultato complessivo è che un osservatore non può sfruttare le fluttuazioni per determinare la propria velocità.

Le fluttuazioni possono però servire per determinare l'accelerazione. Nel 1976 William G. Unruh dell'Università della British Columbia dimostrò che un ipotetico rivelatore di particelle sottoposto a un'accelerazione costante reagirebbe alle fluttuazioni del vuoto come se fosse in quiete in un gas di particelle (e quindi non nel vuoto) con una temperatura proporzionale all'accelerazione. Un rivelatore non accelerato non reagirebbe affatto alle fluttuazioni.

L'idea che la temperatura e l'accelerazione possano essere correlate in questo modo ha condotto a una revisione del concetto di «vuoto» e al riconoscimento dell'esistenza di diversi tipi di vuoto. Uno dei più semplici vuoti non tradizionali si può creare ripetendo, in un contesto quantomeccanico, un esperimento ideale proposto per la prima volta da Einstein. Si immagini un montacarichi chiuso che si sta muovendo liberamente nel vuoto. Uno «spirito scherzoso» si aggrappa a esso, portandolo in uno stato di accelerazione costante con l'estremità superiore in avanti. Si suppone che le pareti del montacarichi siano perfettamente conduttrici, impermeabili alla radiazione elettromagnetica, e che il montacarichi stesso sia completamente vuoto, in modo da non contenere alcuna particella. Einstein introdusse questa descrizione immaginaria per illustrare l'equivalenza tra gravitazione e accelerazione, ma un riesame mostra anche che ci si possono aspettare numerosi effetti strettamente quantomeccanici.

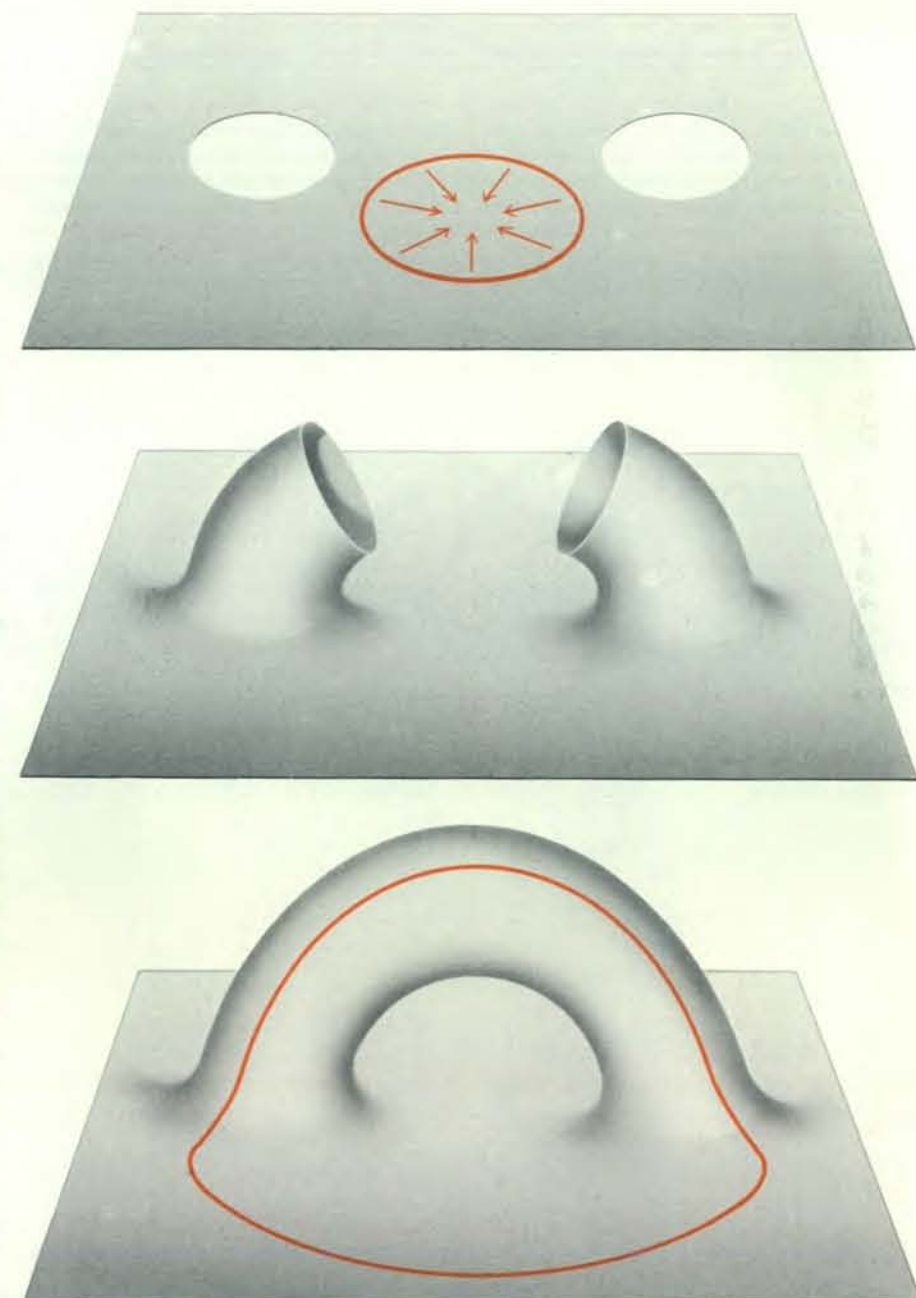
Tanto per cominciare, nell'istante in cui inizia l'accelerazione il pavimento del montacarichi emette un'onda elettromagnetica che si propaga verso il soffitto e rimbalza su e giù. (Il dimostrare perché venga emessa l'onda richiederebbe una dettagliata analisi matematica di un conduttore elettrico accelerato, ma l'effetto è analogo alla creazione dell'onda acustica di compressione che apparirebbe se il montacarichi fosse pieno d'aria.) Se le pareti del montacarichi consentono temporaneamente una certa dissipazione, l'onda elettromagnetica viene trasformata in fotoni con uno spettro energetico termico, o in altre parole in una radiazione di corpo nero caratteristica di una certa temperatura.

Il montacarichi contiene ora un tenue gas di fotoni. Per liberarci dai fotoni possiamo installare un refrigeratore con un radiatore all'esterno, con una certa spesa di energia di fonte esterna. Il risultato finale, quando tutti i fotoni sono stati estratti, è un nuovo vuoto all'interno del montacarichi, un vuoto lievemente diverso dal vuoto standard all'esterno. In primo luogo, infatti, un rivelatore di Unruh che condivide l'accelerazione del montacarichi, e che reagirebbe termicamente alle fluttuazioni del campo se venisse posto nel vuoto standard all'esterno, all'interno non mostra alcuna reazione; in secondo luogo, i due vuoti differiscono per il contenuto di energia.

Per precisare l'energia di un vuoto, è necessario risolvere alcuni problemi delicati della teoria quantistica dei campi. Ho sottolineato prima come un campo libero equivalga a un insieme di oscillatori armonici. Le fluttuazioni dello stato fondamentale degli oscillatori danno al campo nel vuoto un'energia residua, chiamata energia di punto zero. Essendo infinito il numero di oscillatori del campo, sembrerebbe che debba essere infinita anche la densità di energia del vuoto.

Una densità di energia infinita è imbarazzante e i teorici hanno introdotto numerosi dispositivi tecnici per esorcizzarla. Tali dispositivi fanno parte di un programma generale, chiamato teoria della rinormalizzazione, per la trattazione dei vari infiniti che compaiono nella teoria quantistica del campi. Qualsiasi dispositivo adottato deve essere universale, cioè non costruito «su misura» per un particolare problema fisico, ma tale da adattarsi uniformemente a tutti i problemi. Esso deve anche dar luogo a una densità di energia che scompare nel vuoto standard. Quest'ultimo requisito è fondamentale per la coerenza con la teoria di Einstein, perché il vuoto standard è l'equivalente quantistico dello spazio-tempo piatto e vuoto. Se in esso vi fosse energia, esso non sarebbe piatto.

Di regola le varie impostazioni della teoria della rinormalizzazione danno risultati identici quando vengono applicate allo stesso problema, il che dà garanzie sulla loro validità. Quando vengono applicate ai vuoti all'interno e all'esterno del montacarichi, danno una densità di energia nulla all'esterno e una densità di energia negativa all'interno. Un'energia del vuoto negativa costituisce una sorpresa. Che cosa può voler dire meno di niente? Un attimo di riflessione spiega però la ragionevolezza dell'apparente valore negativo. All'interno del montacarichi devono essere aggiunti fotoni termici, perché un rivelatore di Unruh all'interno si comporti come nel vuoto standard all'esterno. Quando si aggiungono i fotoni, la loro energia riporta a zero l'energia totale interna, uguale a quella del vuoto esterno.

Dobbiamo sottolineare che tali strani effetti sarebbero difficili da osservare in pratica. Per le accelerazioni ricorrenti nella vita quotidiana, perfino nelle macchine ad alta velocità, l'energia negativa è di gran lunga troppo piccola per essere rilevata. Esiste però un caso nel quale è stata osservata un'energia negativa del vuoto, almeno indirettamente: in un effetto previsto nel 1948 da H. B. G. Casimir dei Laboratori di ricerca Philips in Olanda. Nell'effetto Casimir vengono affacciate

Un «cunicolo» (*wormhole*) nello spazio-tempo è una struttura ipotetica che potrebbe alterare la topologia dell'universo. In uno spazio piatto un cunicolo si forma praticando due aperture e stirandone i bordi in tubi che vengono poi congiunti. Nel piano originario qualsiasi curva chiusa può essere ridotta a un punto (*in colore*), ma non è possibile per una curva che attraversi il cunicolo. Un cunicolo nello spazio a tre o a quattro dimensioni non è concettualmente differente.

vicinissime nel vuoto due lastre metalliche microscopicamente piane, pulite, parallele e scariche e si vede che si attirano debolmente a vicenda con una forza che si può attribuire a una densità di energia negativa nel vuoto che sta tra di esse.

Il vuoto diventa ancor più complesso quando lo spazio-tempo è curvo. La curvatura influenza la distribuzione spaziale delle fluttuazioni del campo quantistico e, come l'accelerazione, può indurre un'energia del vuoto non nulla. Dal momento che la curvatura può variare da luogo a luogo, può variare anche l'energia del vuoto, mantenendosi positiva in alcuni luoghi e negativa in altri.

In qualsiasi teoria coerente, l'energia si deve conservare. Supponiamo per il momento che un aumento di curvatura provochi un aumento dell'energia del vuoto quantistico. Tale aumento deve venire da qualche parte e, quindi, la stessa esistenza delle fluttuazioni del campo quantistico implica che sia necessaria energia per curvare lo spazio-tempo. Ne consegue che lo spazio-tempo si oppone alla curvatura. È proprio come nella teoria di Einstein.

Nel 1967 il fisico Andrei Sakharov ipotizzò che la gravitazione potesse essere un fenomeno puramente quantistico derivante dall'energia del vuoto e che la costante di Newton G o, in modo equivalente, la rigidità dello spazio-tempo, fosse calcolabile dai principi fondamentali. Quest'idea incontra molte difficoltà. In primo luogo, richiede che la gravità venga sostituita, come campo fondamentale, da qualche «campo di gauge di grande unificazione» suggerito dalle particelle elementari note. Si deve introdurre a questo punto una massa fondamentale per poter ottenere un'altra scala assoluta di unità; quindi una costante fondamentale viene sostituita da un'altra.

In secondo luogo, e forse più importante, la dipendenza calcolata dell'energia del vuoto dalla curvatura conduce a una teoria della gravità più complessa di quella di Einstein. A seconda del numero e del tipo dei campi elementari scelti e del metodo di rinormalizzazione, l'energia del vuoto, anziché aumentare all'aumentare della curvatura, può perfino diminuire. In base a questa relazione di proporzionalità inversa lo spazio-tempo piatto sarebbe instabile e tenderebbe a raggrinzirsi come una prugna. Supporrò qui che il campo gravitazionale sia fondamentale.

Un vero vuoto è definito come uno stato di equilibrio termico alla temperatura dello zero assoluto. Nella gravità quantistica un tale vuoto può esistere soltanto se la curvatura è indipendente dal tempo. Quando la curvatura dipende dal tempo, nel vuoto possono apparire spontaneamente particelle (con il risultato che, ovviamente, non si tratta più di un vuoto).

Il meccanismo di produzione di particelle può essere spiegato anche in termini di oscillatori armonici. Quando cambia la curvatura dello spazio-tempo, cambiano anche le proprietà fisiche degli oscillatori del campo. Supponiamo che un comune oscillatore si trovi inizialmente nel suo stato fondamentale, soggetto a oscillazioni di punto zero. Se una delle sue proprietà, quali la massa o la rigidità della molla, cambia, le sue oscillazioni di punto zero devono a loro volta adattarsi alla variazione. Dopo l'adattamento c'è una probabilità finita che l'oscillatore non si trovi più nel suo stato fondamentale, ma in uno stato eccitato. Il fenomeno è analogo all'aumento di vibrazione indotto in una corda vibrante di un pianoforte quando aumenta la sua tensione; l'effetto è chiamato eccitazione parametrica. Nel campo quantistico, l'analogo dell'eccitazione parametrica è la produzione di particelle.

Le particelle prodotte da una curvatura variabile nel tempo appaiono casualmente. Non è possibile prevedere esattamente in anticipo dove o quando nascerà una data particella. Si può però calcolare la distribuzione statistica dell'energia e della quantità di moto delle particelle. La produzione di particelle è massima dove la curvatura è massima e sta cambiando con la massima rapidità. Essa fu probabilmente molto grande durante il big bang e potrebbe aver avuto un effetto predominante sulla dinamica dell'universo nei primi istanti. Particelle create in questo modo potrebbero spiegare tutta la materia dell'universo.

Tentativi di calcolare la produzione di particelle durante il big bang furono iniziati indipendentemente una decina di anni fa dall'accademico russo Yakov B. Zel'dovich e da Leonard E. Parker dell'Università del Wisconsin a Milwaukee. Da allora, molti altri hanno esaminato il problema, con risultati anche suggestivi, ma senza raggiungere una soluzione definitiva. Inoltre, pende in proposito una domanda fondamentale: quale stato quantico iniziale si deve scegliere per l'istante del big bang? Qui il fisico deve assumere il ruolo di Dio. Nessuna delle proposte finora avanzate sembra perfettamente convincente.

L'altro evento dell'universo durante il quale la curvatura dovrebbe cambiare rapidamente è il collasso di una stella che dà origine a un buco nero. A questo proposito i calcoli quantomeccanici hanno portato a una vera sorpresa, sostanzialmente indipendente dalle condizioni iniziali. Nel 1974 Stephen W. Hawking dell'Università di Cambridge dimostrò che la variazione di curvatura in vicinanza di un buco nero in fase di collasso crea un flusso radiante di particelle. Il flusso è stazionario e continua per lungo tempo dopo che il buco nero è diventato geometricamente quiescente. Esso può continuare poiché pare che il tempo rallenti nell'enorme campo gravitazionale vicino alla superficie di «orizzonte» di un buco nero; per un osservatore esterno tutta l'attività si porta a una condizione virtuale di riposo. Le particelle nate più vicine all'orizzonte vengono maggiormente ritardate nel loro viaggio verso l'esterno.

Anche se il ritardo nell'emissione implica che debba esistere un enorme numero di particelle addensate nei pressi dell'orizzonte, ciascuna «in attesa del proprio turno» per sfuggire, la densità di energia totale in questa regione è effettivamente negativa e piuttosto piccola. L'energia positiva trasportata dalle particelle è ampiamente compensata da un'energia del vuoto enormemente negativa che sarebbe presente se fossero assenti le particelle (per esempio, se il buco nero fosse sempre esistito e non fosse mai stato creato da un collasso gravitazionale).

Si può dimostrare che le particelle emesse sono statisticamente non correlate e che il loro spettro energetico è termico. Il carattere di corpo nero della radiazione è forse la sua proprietà più importante e consente di assegnare a un buco nero sia una temperatura sia un'entropia. L'entropia, che misura il disordine termodinamico del sistema, risulta proporzionale all'area della superficie di orizzonte. Per un buco nero di massa stellare essa è enorme: è di oltre diciannove ordini di grandezza superiore all'entropia della stella collassata che ha formato il buco nero. La temperatura, a sua volta, è inversamente proporzionale alla massa e, se la massa è stellare, è di oltre undici ordini di grandezza inferiore a quella della stella madre.

Dato che la quantità di radiazione emessa da un oggetto dipende dalla sua temperatura, la radiazione di Hawking proveniente da un buco nero astrofisico è del tutto trascurabile. Essa diventa importante solo per buchi neri «mini», ossia quelli di una massa inferiore a 10^{10} grammi. Il solo modo immaginabile in cui si sarebbero potuti formare i minibuchi neri è per compressione durante il big bang. È probabile che allora ne siano stati prodotti in abbondanza, nel qual caso avrebbero contribuito in modo significativo all'entropia dell'universo.

L'energia delle particelle create da una curvatura variabile nel tempo non può essere evocata dal nulla. Essa deriva dallo stesso spazio-tempo. Ne consegue che le particelle agiscono a loro volta sullo spazio-tempo. Sono stati fatti tentativi per calcolare questa «retroazione» nel caso del big bang, per determinare il suo effetto dinamico sull'universo primitivo. Si voleva vedere se la retroazione poteva eliminare la densità iniziale infinita della materia richiesta dalla teoria classica di Einstein e che è un ostacolo per tutte le ricerche ulteriori. Se la si potesse sostituire con una densità semplicemente enorme, ci si potrebbe domandare: che cosa faceva l'universo *prima* del big bang?

Negli anni sessanta Roger Penrose dell'Università di Oxford e Hawking dimostrarono che la teoria classica di Einstein è incompleta. Essa prevede il verificarsi, nel passato o nel futuro, di densità infinite e di curvature infinite sotto una varietà di condizioni attuali fisicamente ragionevoli. Una teoria che prevede un valore infinito per una grandezza osservabile non è più in grado di avanzare previsioni al di là di quel punto. Dato che i fisici credono nella comprensibilità della natura, essi si aspettano che una teoria siffatta richieda un ampliamento fino a comprendere una gamma più ampia di fenomeni. Secondo l'attuale punto di vista prudenziale l'inclusione degli effetti quantistici è la sola terapia ragionevole per l'incompletezza della teoria di Einstein.

Calcoli della retroazione sul big bang eseguiti mediante simulazione numerica su un calcolatore digitale hanno dato finora risultati ambigui. Difficile è stata la determinazione, come dato di ingresso per il calcolatore, di un valore accettabile della densità di energia combinata per le particelle prodotte e per il vuoto quantistico al quale esse sono sovrapposte.

L'effetto della retroazione è di partico-

In linea di principio regioni lontane dell'universo potrebbero essere connesse da un cunicolo, facendo pensare alla possibilità di stabilire tra esse comunicazioni più veloci della luce; in realtà tale schema non può essere valido. Nel cunicolo in alto a sinistra la distanza tra le aperture nel «mondo esterno» è paragonabile alla distanza lungo il «passaggio». Nel cunicolo in basso a sinistra la distanza esterna è molto maggiore. Nei disegni in basso lo spazio rappresentato dal piano appare curvo, ma ciò è solo dovuto al fatto che viene visto dalla prospettiva di uno spazio con un maggior numero di dimensioni; per un osservatore che vive nel piano esso apparirebbe quasi piatto. Che il passaggio sia o no una scorciatoia, è impossibile attraversarlo, dato che un cunicolo collega immancabilmente due buchi neri. Il passaggio «si strozza», come si vede a destra, e qualunque cosa entri viene schiacciata a una densità infinita prima di raggiungere il lato opposto.

La topologia fluttuante, che è un aspetto dello spazio-tempo in alcuni tentativi diretti a formulare una teoria della gravità quantistica, solleva serie difficoltà concettuali. Qui sono mostrate due rappresentazioni di un cunicolo che è appena stato strizzato, lasciando due «increspature». Se un tale evento può aver luogo, dovrebbe essere possibile anche il processo inverso; in altri termini, le increspature dovrebbero essere in grado di fondersi per formare un nuovo cunicolo. Il processo inverso sembra possibile quando le increspature appaiono abbastanza vicine, ma non quando sono molto lontane. I concetti di «vicino» e «lontano» però dipendono dal vedere la superficie immersa in uno spazio con un maggior numero di dimensioni. Per un osservatore che si trovi nello spazio bidimensionale della superficie, gli oggetti rappresentati dai due disegni apparirebbero indistinguibili.

lare importanza nel caso di un buco nero. La radiazione di Hawking «ruba» da un buco nero sia energia sia entropia. La massa del buco di conseguenza diminuisce. Il tasso di diminuzione è dapprima lento, ma cresce rapidamente all'aumentare della temperatura. Alla fine il tasso di variazione diventa tanto elevato che le approssimazioni nei calcoli di Hawking non sono più valide. Non si sa che cosa accada da quel momento in poi. Hawking ritiene che le sue approssimazioni si mantengono qualitativamente corrette e che il buco nero finisca di vivere con un lampo spettacolare, lasciandosi dietro momentaneamente una «singolarità nuda» nella struttura causale dello spazio-tempo.

Qualsiasi singolarità, nuda o meno, costituisce un fallimento della teoria. Se Hawking ha ragione, non è incompleta solo la teoria di Einstein, ma anche la teoria dei quanti. La ragione è dovuta al fatto che, per ogni particella nata al di fuori della superficie di orizzonte, ne nasce una all'interno. Le due particelle sono correlate nel senso che un osservatore potrebbe rilevare «effetti di interferenza di probabilità» se potesse comunicare simultaneamente con entrambe le particelle. Hawking suppone che le particelle all'interno vengano schiacciate fino a una densità infinita e cessino di esistere. Nel momento in cui cessano di esistere viene a cadere la normale interpretazione probabilistica della meccanica quantistica. La probabilità stessa scompare nello schiacciamento a densità infinita.

Secondo un'ipotesi alternativa ed egualmente plausibile la costruzione stessa della teoria quantistica dei campi che viene eretta attorno alla teoria di Einstein impedisce che vengano perse nel collasso sia la probabilità sia l'informazione. È del tutto probabile che la retroazione arrivi a un tale estremo da impedire che lo schiacciamento si estenda all'infinito. L'orizzonte, che è un costrutto matematico e non fisico, potrebbe non essere affatto una barriera strettamente a senso unico. La materia che è collassata formando il buco nero potrebbe alla fine essere spiegata, particella per particella. Nessuno dubita che si abbia un lampo finale di radiazione di Hawking e che si raggiungano densità enormi all'interno del buco. La pressione stessa alla quale sono sottoposte le particelle nucleari potrebbe tuttavia trasformarle in fotoni e in altre particelle prive di massa, che alla fine potrebbero sfuggire, trasportando con sé la poca energia restante e tutte le correlazioni quantistiche. Non è necessario che questi prodotti finali portino via alcuna parte dell'entropia originaria del buco nero. Essa è stata tutta rubata dalla radiazione di Hawking.

Il vuoto quantistico, pensato da John Archibald Wheeler nel 1957, diventa sempre più caotico ispezionando regioni di spazio sempre più piccole. Alla scala del nucleo atomico *(in alto)* lo spazio appare molto regolare. Alle dimensioni di 10^{-30} centimetri *(al centro)* appaiono nella geometria certe asperità. Alla scala della lunghezza di Planck, 1000 volte ancora più piccola *(in basso)*, la curvatura e la topologia dello spazio sono continuamente sottoposte a violente fluttuazioni.

Passo ora alla parte profonda e difficile della gravità quantistica. Quando un effetto quantistico, quale la produzione di particelle o l'energia del vuoto, retroagisce sulla curvatura dello spazio-tempo, la stessa curvatura diventa un oggetto quantistico. Uno schema teorico coerente richiede che lo stesso campo gravitazionale sia quantizzato. Per lunghezze d'onda grandi rispetto alla lunghezza di Planck, le fluttuazioni quantistiche del campo gravitazionale quantizzato sono piccole. Esse possono essere rappresentate con precisione se vengono trattate come una debole perturbazione su un fondo classico. La perturbazione può essere analizzata allo stesso modo di un campo indipendente e contribuisce per la sua parte sia all'energia del vuoto sia alla produzione di particelle.

Alle lunghezze d'onda e alle energie di Planck la situazione è decisamente più complessa. Le particelle associate a un campo gravitazionale debole sono dette gravitoni; sono prive di massa e hanno un momento angolare di spin $2\hbar$. È improbabile che i singoli gravitoni possano mai essere osservati direttamente. La comune materia, perfino un'intera galassia di materia comune, è quasi totalmente trasparente a essi. I gravitoni interagiscono apprezzabilmente con la materia solo quando raggiungono le energie di Planck. Tuttavia, a tali energie, essi sono in grado di indurre curvature di Planck nella geometria di fondo. A questo punto il campo al quale i gravitoni sono associati non è più debole e lo stesso concetto di «particella» è mal definito.

Alle lunghezze d'onda maggiori l'energia trasportata da un gravitone distorce la geometria di fondo. Alle lunghezze d'onda minori essa distorce le onde associate al gravitone stesso. È una conseguenza della non linearità della teoria di Einstein: quando si sovrappongono due campi gravitazionali, il campo risultante non è uguale alla somma dei due componenti. Tutte le teorie di campo significative sono non lineari. Per alcune si può trattare la non linearità con un metodo di approssimazioni successive chiamato teoria delle perturbazioni, nome la cui origine risale alla meccanica celeste. Il metodo si basa sul perfezionamento di un'approssimazione iniziale mediante una serie di correzioni progressivamente più piccole. Quando si applica la teoria delle perturbazioni ai campi quantizzati, essa conduce a infiniti che devono essere eliminati per rinormalizzazione.

Nel caso della gravità quantistica la teoria delle perturbazioni non è applicabile per due motivi. In primo luogo, alle energie di Planck i successivi termini della serie di perturbazioni (cioè le successive correzioni) sono di grandezza analoga. Troncare la serie a un numero finito di termini non porta a una approssimazione valida; si deve invece sommare l'intera serie infinita. In secondo luogo, non possono essere rinormalizzati in modo coerente i singoli termini della serie. In ogni ordine di approssimazione appaiono nuove classi di infiniti, che non hanno alcuna corrispondenza nell'ordinaria teoria quantistica dei campi. Nascono perché nel quantizzare il campo gravitazionale si quantizza lo stesso spazio-tempo. Nell'ordinaria teoria quantistica dei campi lo spazio-tempo è un fondo fisso. Nella gravità quantistica il fondo non solo reagisce alle fluttuazioni quantistiche, ma si suddivide anche tra esse.

Sono stati fatti alcuni tentativi per dare una risposta strettamente tecnica a queste difficoltà sommando sottoinsiemi infiniti di termini della serie di perturbazioni. I risultati, in particolare la completa scomparsa degli infiniti, sono sia suggestivi sia incoraggianti. Dobbiamo però esaminarli con cautela perché per ricavarli sono state operate pesanti approssimazioni e la serie di perturbazioni non viene mai sommata in toto; ciononostante vengono impiegati per calcolare stime migliori dell'effetto della retroazione sul big bang.

Da un punto di vista più ampio ci si deve aspettare l'insorgere di altri problemi la cui soluzione non può neppure essere tentata con la somma della serie. Uno spazio-tempo quantizzato possiede una struttura causale fluttuante e incerta. Alle dimensioni di Planck la stessa distinzione tra passato e futuro diventa nebulosa. Per analogia con l'effetto tunnel nei sistemi atomici, che consente a un elettrone di attraversare una barriera energetica che non potrebbe mai scavalcare, dobbiamo attenderci processi non ammessi nella teoria classica di Einstein, tra i quali viaggi su distanze di Planck a velocità superiori a quella della luce. Non sappiamo affatto come calcolare le probabilità di tali processi. In molti casi non si sa neppure quali siano le domande giuste da porre. Non vi sono esperimenti che possano farci da guida e perciò è ancora lecito indulgere a voli di fantasia.

La dimensionalità dello spazio viene messa in discussione dalla possibilità che lo spazio-tempo abbia una topologia complessa. La superficie mostrata è bidimensionale, ma per le sue connessioni topologiche ha un aspetto tridimensionale. È ammissibile che lo spazio tridimensionale percepito a una scala macroscopica abbia in realtà un minor numero di dimensioni, ma sia topologicamente convoluto.

Uno dei voli di fantasia più ricorrenti, ripetutamente citato nella letteratura sulla gravità quantistica, è il concetto di topologia fluttuante. La nozione fondamentale introdotta da Wheeler nel 1957 è la seguente: le fluttuazioni nel vuoto del campo gravitazionale, come quelle di tutti gli altri campi, aumentano di intensità alle lunghezze d'onda minori. Se si estrapolano alla regione di Planck i risultati standard per il campo debole, le fluttuazioni di curvatura diventano talmente violente da apparire in grado di produrre buchi nello spazio-tempo alterandone la topologia. Wheeler immagina il vuoto in uno stato di perenne agitazione, con la continua comparsa e scomparsa di cunicoli (e di altre strutture più complesse) di dimensioni planckiane. L'agitazione è «visibile» soltanto al livello di Planck. A un livello più grossolano lo spazio-tempo continua ad apparire regolare.

Si può sollevare un'obiezione immediata: ogni variazione di topologia è necessariamente seguita da una singolarità nella struttura causale dello spazio-tempo, cosicché ci si trova di fronte alla stessa difficoltà che nasce nell'interpretazione di Hawking del decadimento dei buchi neri. Supponiamo comunque che l'interpretazione di Wheeler sia corretta. Una delle prime domande alle quali si deve dare risposta è: in quale misura le fluttuazioni topologiche contribuiscono all'energia del vuoto e come influenzano la resistenza alla curvatura dello spazio-tempo (al livello grossolano)? Finora nessuno ha dato una risposta convincente, principalmente perché non è emersa alcuna descrizione coerente del processo stesso di transizione topologica.

Per renderci conto di almeno uno degli ostacoli che si incontrano nella costruzione di tale descrizione, si consideri il processo visualizzato nell'illustrazione a pagina 15. Il disegno fornisce due rappresentazioni dello stesso evento: un cunicolo è stato appena strizzato lasciando due «pseudopodi» residui in uno spazio connesso semplicemente. In una rappresentazione lo spazio si presenta curvo, nell'altra piatto.

Si consideri ora il processo inverso: la formazione di un cunicolo. Se esiste una probabilità finita che un cunicolo scompaia per strizzatura, esiste anche una probabilità finita che se ne formi un altro. A questo punto sorge una nuova difficoltà. Dal punto di vista dell'inversione l'illustrazione rappresenta due pseudopodi cresciuti spontaneamente nel vuoto quantistico. In una rappresentazione appare ragionevole la possibilità che due pseudopodi si uniscano formando un cunicolo. Nell'altra non è così. Eppure la situazione fisica è la stessa nei due disegni. Nel primo caso la formazione del cunicolo sembra ragionevole perché gli pseudopodi appaiono vicini. La «vicinanza», però, non è una proprietà intrinseca della disposizione spaziale, come l'altra rappresentazione mostra chiaramente. Una nozione di «vicinanza» richiede l'esistenza di uno spazio con un maggior numero di dimensioni nel quale sia immerso lo spazio-tempo. Inoltre, lo spazio a più dimensioni deve essere dotato di proprietà fisiche tali che gli pseudopodi possano trasmettere una sensazione di vicinanza gli uni agli altri. Ma allora lo spazio-tempo non è più

Potrebbero esistere altre dimensioni spaziali, oltre a quelle conosciute, se avessero una forma «compatta». Per esempio, una quarta dimensione spaziale potrebbe essere arrotolata su un cilindro con una circonferenza di forse 10^{-32} centimetri. Qui l'ipotetica dimensione compatta è stata «srotolata» e viene rappresentata come l'asse verticale di un diagramma spazio-tempo. La traiettoria di una particella può quindi avere una componente ciclica: ogni volta che raggiunge la massima estensione della dimensione compatta si ritrova al punto di partenza. La traiettoria osservata è la proiezione di quella reale sulle dimensioni macroscopiche dello spazio-tempo. Se la traiettoria è una geodetica, può assumere l'aspetto di una particella elettricamente carica in moto in un campo elettrico. Una teoria di questo genere è stata introdotta negli anni venti da Theodor Kaluza e Oskar Klein, per dimostrare come essa sia in grado di spiegare sia la gravitazione sia l'elettromagnetismo.

l'universo. L'universo è qualcosa di più. Se restiamo fedeli all'idea che le proprietà dello spazio-tempo siano intrinseche e che non siano il risultato di qualcosa di esterno, sembra che una descrizione coerente delle transizioni topologiche, sia fuori dalla nostra portata.

Un'altra difficoltà collegata alle fluttuazioni topologiche è che esse potrebbero danneggiare la dimensionalità macroscopica dello spazio. Se i cunicoli si possono formare spontaneamente, gli stessi cunicoli possono formare altri cunicoli, e così via all'infinito. Lo spazio potrebbe evolversi in una struttura la quale, benché tridimensionale a livello di Planck, presenta quattro o più dimensioni apparenti in una scala maggiore. Un esempio comune di tale processo è la formazione di una schiuma, che è costituita interamente da superfici bidimensionali, ma ha una struttura tridimensionale (*si veda l'illustrazione a pagina 17*).

A causa di difficoltà come queste, alcuni fisici hanno suggerito che la descrizione convenzionale dello spazio-tempo come un continuo regolare perde validità al livello di Planck e deve essere sostituita da qualcos'altro. Non è mai stato molto chiaro di che cosa possa consistere questo qualcos'altro. Visto il successo della descrizione del continuo su scale di lunghezza che coprono oltre quaranta ordini di grandezza (o sessanta se si suppone che la possibile perdita di validità intervenga solo al livello di Planck), parrebbe ugualmente ragionevole che la descrizione del continuo sia valida a tutti i livelli e che le transizioni topologiche semplicemente non esistano.

Anche se la topologia dello spazio è immutabile, non è necessariamente semplice, neppure a un livello microscopico. È concepibile che lo spazio possa avere una struttura a schiuma sin dall'inizio, nel qual caso la sua apparente dimensionalità potrebbe essere maggiore della sua dimensionalità reale. La sua dimensionalità apparente, però, potrebbe anche essere inferiore alla sua dimensionalità reale.

Quest'ultima possibilità è stata proposta in una teoria avanzata da Theodor Kaluza nel 1921 e da Oskar Klein nel 1926. Nella teoria di Kaluza-Klein lo spazio è tetradimensionale e lo spazio-tempo è pentadimensionale. La ragione per la quale lo spazio *appare* tridimensionale è che una delle sue dimensioni è cilindrica, come nell'universo discusso in precedenza, ma con una differenza importante: la circonferenza dell'universo nella direzione cilindrica, invece di essere di miliardi di anni luce, è soltanto di poche (forse 10 o 100) unità di Planck. Di conseguenza, un osservatore che tenti di penetrare nella quarta dimensione spaziale si ritrova quasi istantaneamente al punto di partenza. In realtà ha poco significato parlare di tale tentativo, perché gli atomi stessi dei quali l'osservatore è composto sono molto più grandi della circonferenza del cilindro. La quarta dimensione è semplicemente non osservabile come tale.

Ciononostante, essa può manifestarsi in un altro modo: come luce! Kaluza e Klein dimostrarono che se il loro spazio-tempo pentadimensionale viene trattato matematicamente esattamente allo stesso modo in cui lo spazio-tempo tetradimensionale viene trattato da Einstein, la loro teoria è equivalente alla teoria dell'elettromagnetismo di Maxwell combinata con la teoria della gravitazione di Einstein. Le componenti del campo elettromagnetico sono implicite nell'equazione soddisfatta dalla curvatura dello spazio-tempo. Kaluza e Klein inventarono in tal modo la prima valida teoria unificata dei campi, una teoria che forniva una spiegazione geometrica della radiazione elettromagnetica.

In un certo senso la teoria di Kaluza-Klein ebbe fin troppo successo. Pur fondendo le teorie di Maxwell e di Einstein, essa non prevedeva nulla di nuovo e quindi non poteva essere confrontata con le altre teorie. Il motivo fu che Kaluza e Klein avevano imposto limiti al modo in cui lo spazio-tempo può curvarsi nella dimensione aggiuntiva. Se fossero stati rimossi tali limiti, la teoria avrebbe previsto nuovi effetti, ma non sembrava che gli effetti trovassero corrispondenza nella realtà. La teoria quindi venne considerata per molti anni come una piacevole curiosità e relegata in un angolo.

La teoria di Kaluza-Klein venne tirata fuori dall'angolo negli anni sessanta quando si scoprì che le nuove teorie di gauge che stavano destando un interesse crescente si potevano riformulare come teorie di Kaluza-Klein nelle quali lo spazio è dotato non di una sola, ma di più dimensioni microscopiche aggiuntive. Si cominciò a scoprire che tutta la fisica si poteva spiegare in termini geometrici. Divenne importante a questo punto domandarsi che cosa accade se si rimuovono i limiti alla curvatura nelle dimensioni compatte.

Accade che nelle dimensioni aggiuntive si prevedono fluttuazioni di curvatura, che si manifestano come particelle di grande massa. Se la circonferenza delle dimensioni aggiuntive è di 10 unità di Planck, la massa delle particelle associate è circa un decimo della massa di Planck, pari a circa un microgrammo. Dato che l'energia necessaria per creare tali particelle è enorme, esse non vengono quasi mai prodotte. Pertanto, che si impongano o meno limiti alle fluttuazioni di curvatura fa poca differenza. I problemi rimangono. Quello principale è che la curvatura estrema delle dimensioni aggiuntive dà origine a una grandissima densità di energia nel vuoto classico. L'ipotesi di una grande energia del vuoto è da scartarsi in base alle osservazioni.

I modelli di Kaluza-Klein non sono mai stati seguiti con molta attenzione e il loro ruolo in fisica è ancora incerto. Però, negli ultimi due o tre anni essi sono stati profondamente riesaminati, questa volta in connessione con la generalizzazione della teoria di Einstein chiamata supergravità, ideata nel 1976 da Daniel Z. Freedman, Peter van Nieuwenhuizen e Sergio Ferrara e (in una versione migliorata) da Stanley Deser e Bruno Zumino.

Una delle inadeguatezze dei modelli standard di Kaluza-Klein consiste nel fatto che prevedono soltanto l'esistenza di particelle con momenti angolari di spin 0, $\hbar$ e $2\hbar$, e anche tali particelle sono o prive di massa o superpesanti. Non c'è posto per le particelle della materia comune, la maggior parte delle quali ha momento angolare di spin $1/2\hbar$. Si scopre che, se la teoria di Einstein viene sostituita dalla supergravità e se si tratta lo spazio-tempo col metodo di Kaluza-Klein, si ottiene una vera unione di tutte le varietà di spin.

Nel «supermodello» di Kaluza-Klein oggi più diffuso vengono attribuite allo spazio-tempo altre sette dimensioni. Queste dimensioni hanno la topologia di una ettasfera, uno spazio che possiede alcune proprietà affascinanti. La teoria che ne risulta è straordinariamente ricca e complessa, dal momento che prevede enormi supermultipletti di particelle. Le masse delle particelle sono ancora o nulle o estremamente grandi, ma è possibile che una «rottura» della simmetria della ettasfera attribuisca ad alcune particelle una massa più realistica. Rimane ancora la grande energia del vuoto classico, ma potrebbe essere annullata da un'energia negativa del vuoto quantistico. Non si sa ancora se queste strategie per modificare la teoria avranno successo.

Se Einstein potesse ritornare a verificare che cosa è stato della sua teoria, ne sarebbe certamente stupefatto e, credo, compiaciuto. Sarebbe compiaciuto che i fisici, dopo anni di esitazione, abbiano finito con l'accettare il suo punto di vista secondo il quale le teorie matematicamente eleganti meritano di essere studiate anche se non sembrano corrispondere immediatamente alla realtà. Sarebbe anche compiaciuto che i fisici oggi osino sperare di ricavare una teoria unificata dei campi. E sarebbe infine particolarmente soddisfatto di scoprire che il suo vecchio sogno che tutta la fisica si possa spiegare in termini geometrici pare si stia avverando.

Soprattutto, però, sarebbe meravigliato. Meravigliato del fatto che la teoria dei quanti abbia ancora un ruolo di preminenza nella sua forma originale su tutte le altre teorie, arricchendo la teoria dei campi e venendo a sua volta da essa arricchita. Einstein non credette mai che la teoria dei quanti rappresentasse la verità finale. Egli non accettò mai l'indeterminismo che essa implica ed era convinto che sarebbe stata un giorno sostituita da una teoria dei campi non lineare. È accaduto esattamente il contrario. La teoria dei quanti ha invaso la teoria di Einstein e l'ha trasformata.

NOVITÀ NELLA SERIE LE SCIENZE quaderni

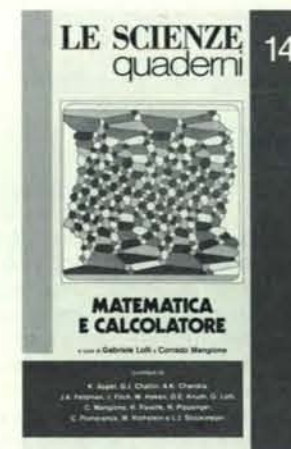
n. 13 febbraio 1984

Un'approfondita ed esauriente descrizione dei processi orogenetici che hanno dato origine ai principali sistemi di catene montuose della Terra.

In questo numero:

Teorie orogenetiche prima della «tettonica a zolle» di M. Parotto
Geosinclinali, orogenesi e crescita dei continenti di R. S. Dietz
La subduzione della litosfera di M. N. Toksöz
Il ruolo dei processi di subduzione di R. Funiciello
La crescita del Nord America di D. L. Jones, A. Cox, P. Coney e M. Beck
Gli Appalachi meridionali di F. A. Cook, L. D. Brown e J. E. Oliver
L'evoluzione delle Ande di D. E. James
La collisione tra India ed Eurasia di P. Molnar e P. Tapponier
Evoluzione e struttura delle Alpi di H. P. Laubscher
Struttura profonda dell'area mediterranea di G. F. Panza, G. Calcagnile, P. Scandone e S. Mueller

n. 14 marzo 1984

Per la prima volta proposta a livello divulgativo un'ampia analisi dell'influsso che la rivoluzione informatica ha avuto sulle scienze matematiche.

In questo numero:

Matematica e calcolatore di G. Lolli
Gli algoritmi di D. E. Knuth
Definizioni di algoritmo di G. Lolli
Linguaggi di programmazione di J. A. Feldman
Algebra e calcolatore di R. Pavelle, M. Rothstein e J. Fitch
Alla ricerca dei numeri primi di C. Pomerance
La teoria della complessità di N. Pippenger
Problemi intrinsecamente difficili di L. J. Stockmeyer e A. K. Chandra
Il problema dei quattro colori di K. Appel e W. Haken
Casualità e dimostrazione matematica di G. J. Chaitin

Otto quaderni all'anno, ogni mese da ottobre a maggio.
In vendita in edicola e in libreria.
Prezzo di copertina L. 4.500

Alternanza di prede in un ecosistema semplice

Nell'isola di Terranova, abitata da poche specie di mammiferi, le linci hanno predato le lepri finché queste hanno subito un collasso numerico; da allora si sono rivolte a giovani caribù, ma il ciclo continua

di Arthur T. Bergerud

In moltissime parti del mondo, la catena ecologica che consente il trasferimento di materia e di energia dai semplici organismi fotosintetizzanti ai mammiferi carnivori, per il tramite degli erbivori, è assai complicata. Ogni piccola zona ospita generalmente molte specie di prede e di predatori; per ogni predatore vi sono molti tipi di prede e per ogni preda molti predatori. Tuttavia, in alcune regioni, le condizioni geografiche hanno portato alla formazione di un sistema ecologico relativamente semplice, comprendente solo poche specie animali. Ecosistemi di questo tipo possono costituire un campo d'indagine per lo studio dei principi che regolano la consistenza delle popolazioni animali. Uno di essi si trova sulla vasta isola di Terranova, dove vivono solo 14 specie indigene di mammiferi, nove delle quali sono carnivore. In una zona in cui i carnivori sono così numerosi e gli erbivori di cui nutrirsi talmente scarsi, qualsiasi oscillazione delle specie predate può avere sensibili ripercussioni, anche per l'assenza di un'attenuazione degli effetti che si manifesterebbe senz'altro in un ecosistema più complesso. Un effetto del genere è stato osservato a Terranova agli inizi del secolo, quando il caribù e la lepre artica, due fra gli animali più predati, cominciarono a diminuire di numero. Si è potuto stabilire che in entrambi i casi la causa di questo arretramento era da imputare al rapido incremento numerico delle linci.

Per molti secoli le linci erano state rare a Terranova, ma fra il 1860 e il 1870 venne introdotta nell'isola la lepre scarpa da neve (*Lepus americanus*) e, grazie a questa nuova fonte alimentare, la loro popolazione si sviluppò rapidamente. Agli inizi di questo secolo tuttavia le lepri avevano raggiunto la soglia critica e subirono un «collasso», cioè diminuirono in breve tempo, privando le linci della principale voce della loro dieta. Di conseguenza, questi predatori pieni di risorse cominciarono ad attaccare i giovani esemplari di caribù e l'altra specie di lepre esistente sull'isola, la lepre artica. Quando la popolazione delle lepri scarpa da neve aumentò nuovamente, le linci ricominciarono a predarle.

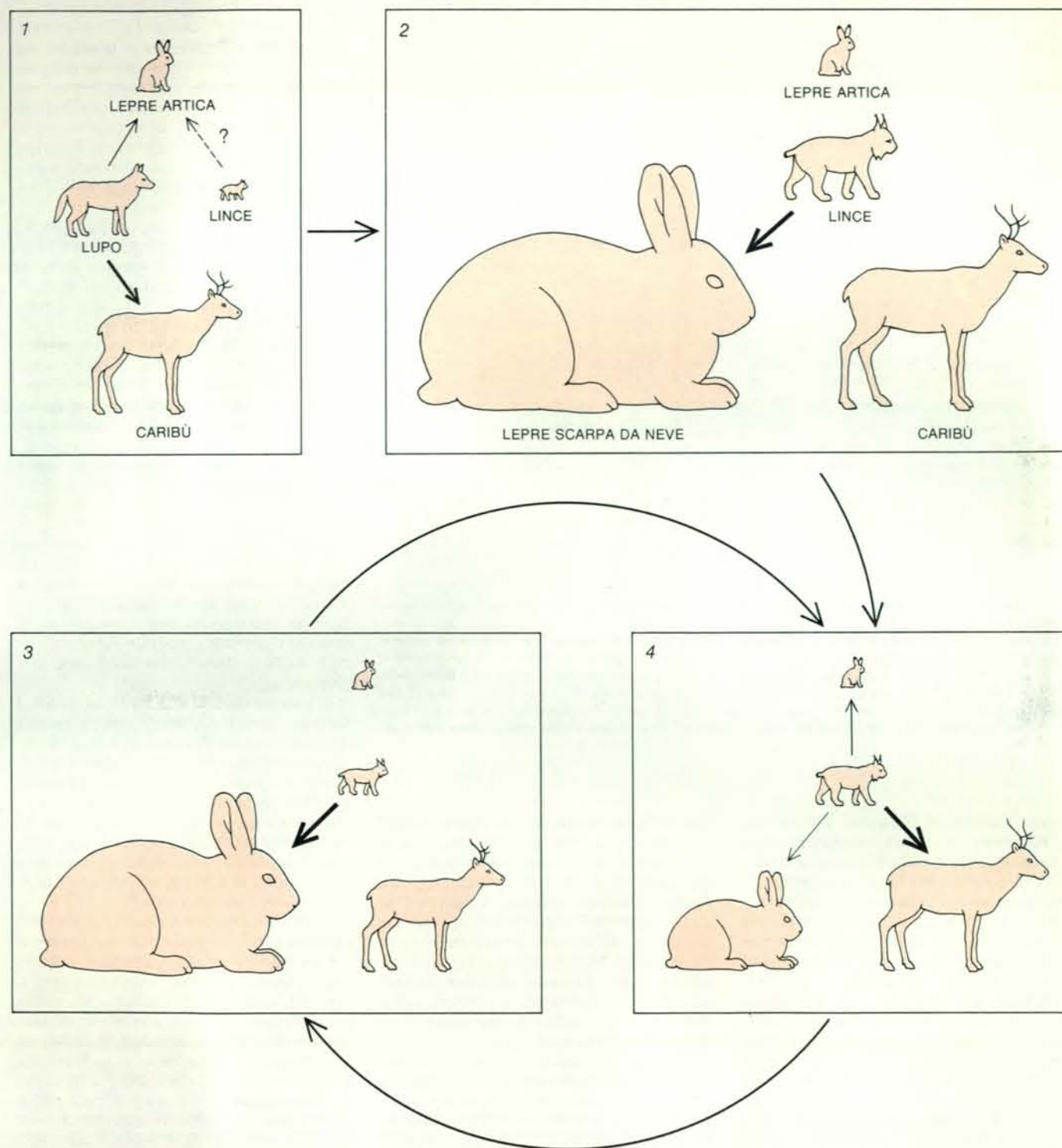
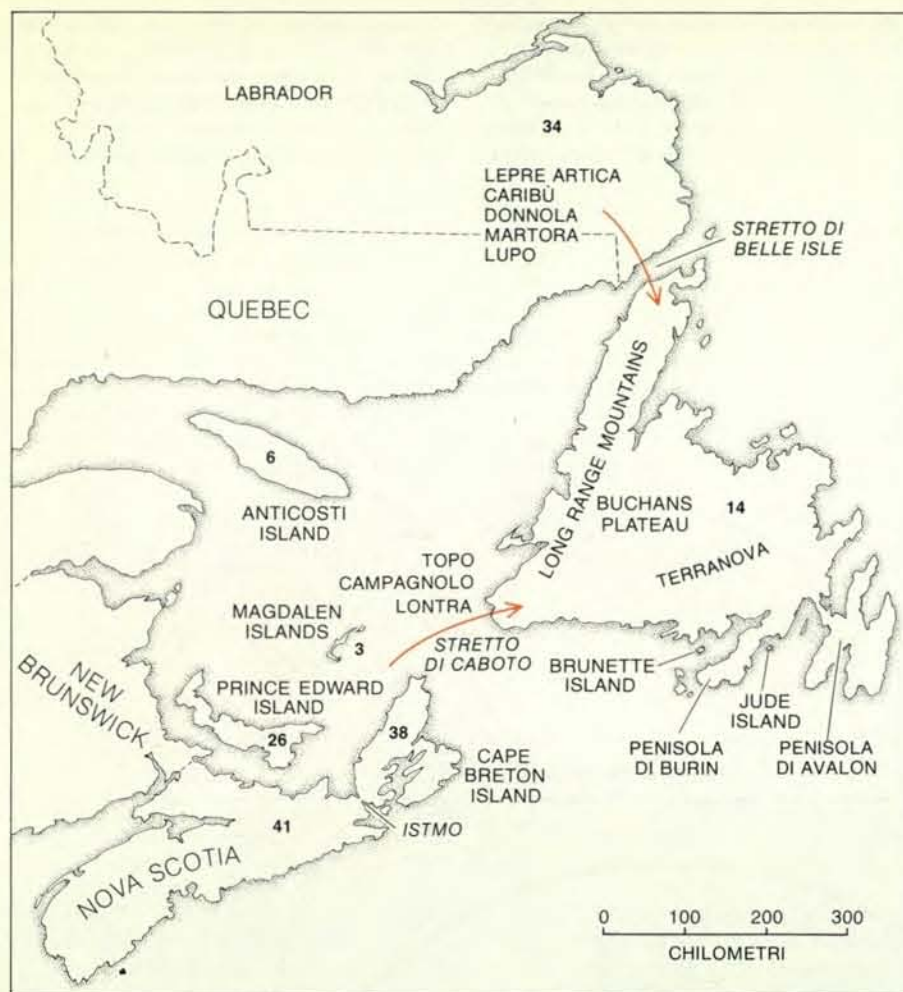
Il ciclo si è ripetuto a più riprese e così, negli ultimi anni, la lince, la lepre scarpa da neve e il caribù hanno fatto parte di un triangolo in cui le popolazioni delle due specie predate hanno oscillato, aumentando e diminuendo a cicli alterni. La lince limita anche il numero delle lepri artiche, un tempo numerose a Terranova, ma ora scarse. L'introduzione di una singola nuova specie all'interno di questo ecosistema semplice ha dato origine a una lunga, intricata matassa di conseguenze per quanto riguarda l'equilibrio del sistema. Se la si riuscisse a dipanare, essa chiarirebbe le trame più fini della dinamica della popolazione animale.

Il sistema ecologico di Terranova è nettamente distinto da quello delle contigue regioni continentali canadesi. Molti fattori hanno contribuito a questo isolamento ed è dimostrato che quando, circa 18 000 anni or sono, avanzò la glaciazione del Wisconsin, il ghiaccio non ricoprì completamente l'isola. Una parte di questa potrebbe aver costituito un rifugio che avrebbe consentito la sopravvivenza dei mammiferi locali. Alcune specie autoctone si sono dunque conservate e hanno continuato a evolversi indipendentemente dai mammiferi del continente. Nove delle 14 specie indigene di mammiferi di Terranova sono sufficientemente differenziate da quelle continentali da poter essere classificate come sottospecie, a dimostrazione di migliaia di anni d'isolamento genetico.

Alcune specie, in particolare i predatori, potrebbero aver raggiunto l'isola su banchi di ghiaccio galleggianti o su ponti di ghiaccio durante il periodo postglaciale. Tuttavia, ai fini di uno scambio biologico completo fra l'isola e le regioni costiere del Canada, il mare ha continuato a funzionare da barriera. Nel Labrador, che dista soli 17,6 chilometri da Terranova, attraverso lo stretto di Belle Isle, vi sono 34 specie di mammiferi. Sulla Cape Breton Island, 112 chilometri a sud-ovest di Terranova, ma collegata con la terraferma, le specie sono 38.

Il mare che circonda Terranova ha agito non solamente da barriera, riducendo il numero globale delle specie, ma anche da filtro, selezionando determinati tipi di mammiferi e impedendo ad altri di raggiungere l'isola. Era prevedibile che questi avventurosi predatori riuscissero meglio delle loro prede a colonizzare l'isola: in origine questa ha ospitato sette quadrupedi carnivori, il lupo comune (*Canis lupus*), la volpe rossa (*Vulpes vulpes*), la lince comune (*Lynx lynx*), la lontra canadese (*Lutra canadensis*), l'orso nero americano o baribal (*Ursus americanus*), l'ermellino (*Mustela erminea*) e la martora americana (*Martes americana*). I cinque erbivori di cui si nutrivano questi predatori, prima dell'arrivo della lepre scarpa da neve, erano il castoro canadese (*Castor canadensis*), il topo campagnolo della Pennsylvania (*Microtus pennsylvanicus*), il caribù (*Rangifer tarandus*), la lepre artica (*Lepus arcticus*) e il topo muschiato (*Ondatra zibethica*). Vi erano anche due specie di chirotteri.

In questo ordinamento, l'abituale piramide delle specie è invertita. Dato che i sistemi biologici, come tutti gli altri sistemi in cui vi sia trasporto di energia, hanno un rendimento limitato, la massa totale dei predatori è molto inferiore alla massa totale delle prede. In base allo stesso principio, il numero delle specie predatrici è generalmente minore del numero delle specie predate. Le faune invertite semplici, come quella di Terranova, si dimostrano fragili, in parte perché costituiscono un'eccezione alle regole che solitamente governano le catene alimentari. Il mio lavoro a Terranova sta a dimostrare quanto possa essere vulnerabile un simile sistema. Dal 1956 al 1967 ho prestato servizio come biologo distrettuale per conto della Newfoundland Division of Wildlife e il mio compito principale era quello di individuare la causa di una brusca diminuzione degli esemplari di caribù. Il risultato delle mie ricerche vale anche per altre specie.

Nel 1900, quando la regione interna di Terranova era abitata esclusivamente dalla piccola tribù indiana dei Micmac, i branchi di caribù erano assai consistenti. Alla fine del XIX secolo, i naturalisti A. A. Radclyffe Dugmore e J. G. Millais avevano calcolato il numero complessivo di questi animali ed erano arrivati a una cifra fra 150 000 e 200 000 esemplari. Queste valutazioni erano però semplici congetture più o meno qualificate; Dugmore e Millais, infatti, non si erano preoccupati di contare o campionare sistemati-

Rappresentazione schematica del ciclo di alternanza di prede a Terranova. Per millenni, fino alla metà del XIX secolo, il lupo è sempre stato il principale predatore del caribù, essendo scarso il numero delle linci indigene (*1*). Le lepri scarpa da neve, introdotte nel 1864 a scopi alimentari, si sono poi moltiplicate in breve tempo, raggiungendo il loro massimo sviluppo verso il 1900 (*2*). Anche le linci, che si nutrivano di esse, si sono diffuse rapidamente, mentre il lupo si è estinto verso il 1911. Attorno al 1915, la popolazione di lepri scarpa da neve ha subito un crollo e le linci hanno cominciato a predare i piccoli di caribù e le lepri artiche, riducendo considerevolmente il numero di queste due specie predate (*3*). Nel frattempo, le lepri scarpa da neve avevano dato inizio a un proprio ciclo decennale di crescita e diminuzione della popolazione. Quando la popolazione di lepri era al culmine, la lince riprendeva a predare, consentendo una rinnovata crescita numerica dei caribù (*4*). Per decenni la lince ha alternato la predazione del caribù e quella della lepre scarpa da neve, seguendo cicli di 10 anni. Attualmente le lepri artiche sono divenute assai rare e vivono per la maggior parte in zone di altopiano; pertanto la loro popolazione è meno fluttuante delle altre.

La fauna originaria di Terranova è circoscritta a 14 specie di mammiferi, mentre le regioni canadesi contigue ne ospitano un numero maggiore. Cinque specie possono essere pervenute a Terranova dal Labrador su banchi galleggianti o ponti di ghiaccio; due provengono forse dalle Maritime Provinces. Il mare ha agito da barriera per le specie e da filtro per i tipi di animali che hanno raggiunto l'isola. Di conseguenza la piramide alimentare delle specie che vivono a Terranova risulta capovolta: delle suddette 14 specie, nove sono carnivore e solo cinque sono erbivore.

camente i branchi di caribù. Basandomi su colloqui con vecchi cacciatori, sulla documentazione di abbattimenti di questi animali e sulle loro attuali presenze, ho calcolato che nel 1900 la popolazione dei caribù a Terranova doveva essere di 40 000 capi, un numero rispettabile per un'isola che ha una superficie di soli 100 000 chilometri quadrati.

Appena 25 anni dopo i caribù erano quasi del tutto estinti e, verso il 1925, Dugmore riteneva che gli animali sopravvissuti fossero solo 200. Questa considerevole diminuzione si verificò anche se verso il 1911 il lupo, che era stato per millenni il principale predatore del caribù, si era estinto.

Verso la fine dell'Ottocento e agli inizi del Novecento era diminuita, rispetto alla quantità iniziale, anche la popolazione autoctona di lepri artiche, quantunque non esista, sul numero originario di questi animali e sul loro tasso di estinzione, una documentazione altrettanto buona di quella riguardante il caribù. Si può solamente presumere il numero di lepri artiche indigene esistenti prima dei grandi insediamenti umani a Terranova, ma la maggioranza degli osservatori concorda nel dire che il territorio frequentato un tempo dalla lepre artica si estendeva lungo la Northern Peninsula fino al Buchans Plateau e al settore meridionale delle Long Range Mountains; da qui volgeva a oriente lungo la costa meridionale fino alle penisole di Burin e di Avalon, comprendendo in pratica buona parte della superficie dell'isola.

L'indizio più evidente di un drastico calo numerico della popolazione di lepri artiche è la costante riduzione della loro area di distribuzione a partire dal 1900. Outram Bang, un naturalista che ha descritto sette delle 14 specie autoctone di Terranova, scriveva nel 1913: «A Terranova la lepre artica è ormai molto rara e localizzata, e si trova solo sulle cime delle alte montagne.» Nel corso degli anni cinquanta, la specie era circoscritta agli altipiani delle Long Range Mountains settentrionali e meridionali e al Buchans Plateau e, lungo la costa meridionale, si spingeva fino alla Bay d'Espoir, ma anche in quest'area era molto scarsa. La popolazione totale durante gli ultimi decenni è stata probabilmente inferiore ai 1000 esemplari.

Quando sono arrivato sull'isola, ho trovato che i caribù presentavano indici di riduzione molto maggiori delle lepri artiche. L'indagine sulla diminuzione numerica dei caribù si è concentrata inizialmente sui fattori che incidono sul tasso di mortalità degli individui più giovani: in molte specie di mammiferi le classi più giovani sono l'anello più debole della sopravvivenza di una popolazione e di fatto, la riduzione nei branchi di caribù sembrava doversi imputare in gran parte a un incremento del tasso di mortalità fra gli esemplari in tenera età.

In un branco di caribù, i maschi adulti pascolano per la maggior parte dell'anno separati dalle femmine; verso la metà di ottobre gli individui dei due sessi si riuniscono per il periodo nuziale, che si protrae per circa 10 giorni. In questo frattempo circa l'80 per cento delle femmine viene ingravidato; le rimanenti lo sono nel corso di un periodo di ricettività sessuale che si verifica circa 10 giorni dopo.

Il periodo di gestazione del caribù è di circa 229 giorni e, per effetto della contemporaneità degli accoppiamenti, quasi tutti i piccoli nascono entro un arco di 14 giorni, che ha inizio attorno al 24 maggio. Le femmine partoriscono in grossi gruppi in luoghi prefissati e generalmente ritornano ogni anno nello stesso posto. La maggior parte di queste «zone di parto» a Terranova sono aperte, paludose, circondate da una vegetazione cespugliosa. Al termine della stagione delle nascite, alcuni branchi si disperdono nei boschi per sfuggire agli sciami di insetti.

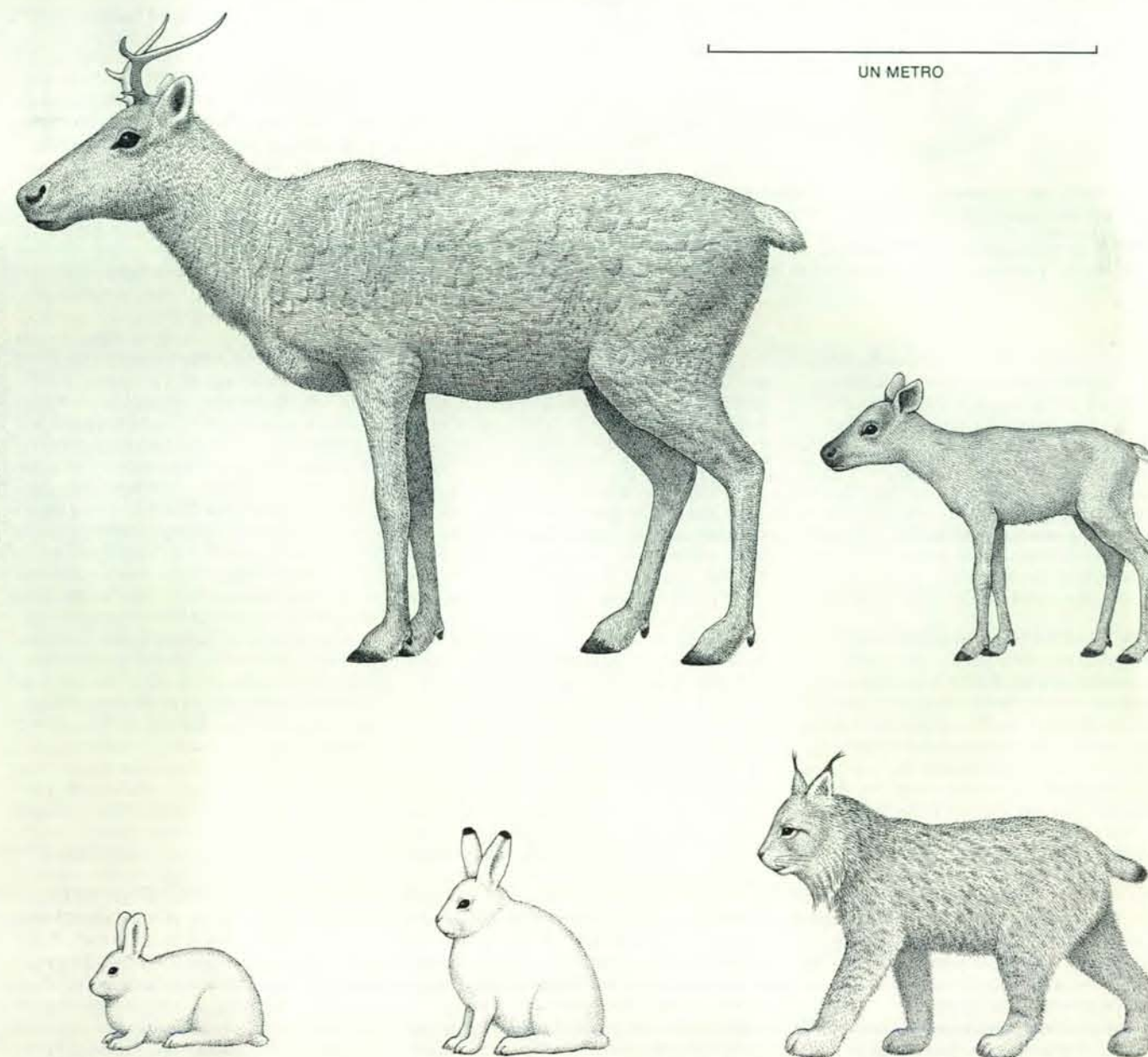
Ogni primavera, dal 1957 al 1964, i biologi hanno rinvenuto molti piccoli morti nelle zone di parto: i resti si trovavano facilmente poiché la madre è solita rimanere presso il piccolo per parecchi giorni dopo la morte. Molti cadaveri presentavano ascessi sul collo provocati da un batterio patogeno comune: *Pasteurella multocida*, del tipo mucoide A. La causa immediata della morte era la setticemia, conseguente all'infezione.

Dato che le femmine partoriscono secondo un ciclo annuale ben sincronizzato e la percentuale delle partorienti è abbastanza costante di anno in anno, non è difficile calcolare il numero di nuovi membri che si aggiunge annualmente a un branco di caribù in assenza di fattori di perturbazione. Sulla base di questi calcoli è risultato evidente che il tasso di mortalità fra i giovani caribù era piuttosto alto e che i piccoli di cui si ritrovavano i resti costituivano solo una minoranza degli esemplari morti. Dal 1958 al 1967, il 27 per cento in media dei piccoli scomparve entro due settimane dalla nascita. Conteggiando le femmine accompagnate dai piccoli e quelle che non lo erano, si è potuto stimare che, a ottobre, circa il 70 per cento delle madri aveva perso i propri figli. Dopo ottobre, invece, il tasso di mortalità calava bruscamente e la maggioranza dei piccoli sopravvissuti fino a quel mese riusciva poi a superare l'inverno. Il fattore limitante le dimensioni dei branchi era quindi la mortalità nel corso dei primi quattro mesi di vita.

Oltre alla scomparsa di giovani esemplari e alla concentrazione dei decessi nel periodo primaverile, si presentavano altri interrogativi. Innanzitutto, al momento della nascita il rapporto fra maschi e femmine era di 52 a 48, mentre entro l'autunno i termini si invertivano e vi erano 62 femmine contro 38 maschi. In secondo luogo, la mortalità globale era più accentuata nelle zone di parto situate alle basse quote: era questa una constatazione inattesa poiché in queste zone il fattore freddo era al minimo. Nelle zone alpine, con clima notevolmente più rigido, il tasso di mortalità risultava inferiore. Il fatto più problematico era che, quando si contavano i branchi, cominciava a delinearsi uno schema ciclico. I computi annuali dimostravano che la percentuale di piccoli sopravvissuti era aumentata nel 1958 e 1959, per poi scemare costantemente dal 1960 al 1963.

Le tessere del mosaico hanno cominciato a comporsi nel 1964 quando è stata individuata la causa degli ascessi. Nella primavera di quell'anno fu trovato un piccolo caribù morto, con quattro punture sul collo, ma senza ascessi. Sembrava probabile che anche gli altri piccoli trovati morti avessero le stesse lesioni, ma che queste fossero state celate dagli ascessi.

La ristretta fauna di Terranova ha semplificato la ricerca della causa di queste lesioni. Fra i predatori locali, solo la lince aveva i denti canini della giusta misura e disposizione per causare quelle quattro morsicature regolarmente spaziate. Ci siamo procurati un teschio di lince e abbiamo potuto verificare che i canini corrispondevano perfettamente alle ferite.

Abbiamo effettuato una coltura di saliva di lince e abbiamo constatato che conteneva *Pasteurella multocida*. Successivamente abbiamo chiuso in un recinto tre piccoli di caribù con una lince e abbiamo osservato che questa li azzannava a turno sul collo. Abbiamo impedito alla lince di ucciderli, togliendoli dal recinto. In breve tempo sul loro collo si sono sviluppati

Qui sono illustrate quattro specie che partecipano al ciclo di alternanza di prede a Terranova. Il caribù (*Rangifer tarandus*) ha un ciclo riproduttivo sincrono e i piccoli vengono partoriti in maggio e giugno, in un periodo di due settimane, in tradizionali zone di parto. Negli anni cinquanta e sessanta, in tali zone sono stati rinvenuti molti cadaveri di piccoli caribù che presentavano ascessi sul collo. Le difese primarie della lepre artica, *Lepus arcticus* (*in basso al centro*), e della lepre scarpa da neve, *Lepus americanus* (*in basso a sinistra*), sono la vigilanza e la velocità. La seconda specie frequenta la foresta boreale, dove la neve è soffice; la prima, invece, vive nella tundra ventosa dove la neve è gelata. Qui appaiono nel loro aspetto invernale. La lince (*Lynx lynx*), attacca la preda scattando dal sottobosco e l'afferra per il collo.

Nel 1964 è stata scoperta l'origine degli ascessi riscontrati sui giovani caribù morti. Venne rinvenuta una carcassa con quattro ferite in forma di foro sul collo, ma priva di ascessi. Quando fu trovato un cranio di lince, si vide che i suoi quattro canini corrispondevano alle ferite. Una coltura di saliva di lince ha prodotto i batteri patogeni precedentemente individuati negli ascessi. La mortalità fra i giovani caribù era dunque da imputarsi alla predazione da parte delle linci.

ascessi per infezione da *Pasteurella* ed essi sono morti rispettivamente dopo quattro, cinque e 15 giorni dalla morsicatura.

Questa indagine ha suggerito che la causa della riduzione dei branchi di caribù risiedeva nella predazione da parte delle linci. I piccoli morti che avevamo trovato sul campo erano riusciti a sfuggire dopo essere stati attaccati, per morire poi di setticemia. I biologi hanno esplorato le zone boschive ai margini delle zone di parto, scoprendo i resti di altri giovani caribù.

Come altri felini predatori, le linci afferrano spesso la preda per il collo e la trascinano al riparo della vegetazione. Evidentemente la lince sta in agguato nel sottobosco, presso la zona di parto, in attesa che qualche piccolo caribù si avvicini imprudentemente, poi balza fuori, lo azzanna per il collo e lo trascina nella macchia. I maschi sono più esposti delle femmine a questi attacchi, poiché sono più intraprendenti e spesso si allontanano dalla madre per esplorare il limitare del bosco: le femmine, invece, preferiscono rimanere presso la madre. Le femmine adulte di Terranova misurano circa 1,20 metri al garrese, pesano circa 100 chili e la loro mole è più che sufficiente a sventare l'attacco di una lince che pesa attorno ai 10 chili.

Per verificare l'ipotesi secondo cui le linci sarebbero state responsabili del calo numerico dei branchi di caribù, abbiamo teso trappole attorno a due zone di parto, allontanandole così da queste zone. È stato così possibile rilevare un aumento statisticamente significativo di quella parte di giovani caribù che riusciva a sopravvivere e ad aggiungersi al branco. Abbiamo poi introdotto dei caribù in molte piccole isole lungo le coste di Terranova, dove non esistevano linci, e - ai fini di un controllo sperimentale - anche in habitat di terraferma dove i caribù locali si erano estinti mentre le linci erano ancora presenti. Su quelle piccole isole i caribù aumentarono fino a raggiungere quasi il livello massimo della specie, mentre negli habitat frequentati dalle linci l'aumento è stato pari a circa la metà. Durante questo stesso periodo la consistenza del branco principale di Terranova ha subìto lievissime variazioni: comprendeva 6100 capi nel 1961 e 6200 nel 1966.

La Jude Island era la località in cui i caribù erano rapidamente aumentati, dopo la loro introduzione da parte del Newfoundland Wildlife Service, diretto da Eugene Mercer. Mercer si è particolarmente impegnato nel lavoro sull'isola: apparecchi radiotrasmittenti sono stati collegati a due linci, poi liberate per poterne osservare gli effetti sul branco in un ambiente non frequentato, fino a quel momento, da altri predatori. Nel corso della prima stagione le linci hanno ucciso circa il 65 per cento dei piccoli caribù, mentre durante la stagione precedente tutti i nuovi nati erano sopravvissuti.

Il lavoro compiuto alla Jude Island ha anche chiarito una delle ragioni per cui la madre si trattiene per diversi giorni presso il corpo del suo piccolo. I movimenti delle linci, segnalati dal trasmettitore, hanno permesso di stabilire che, dopo essere stata distolta dal suo attacco dalla madre, la lince rimane spesso nascosta nella macchia circostante in attesa che quella si allontani. Quando le linci sono state allontanate dall'isola il tasso di sopravvivenza dei piccoli caribù è subito risalito al precedente elevato livello.

Dopo l'esperimento della Jude Island, anche gli osservatori più scettici si sono convinti che a Terranova la lince, che non era mai stata segnalata come un pericoloso predatore di caribù sul continente canadese, costituiva il principale fattore limitante della popolazione di questo cervide. Il lupo, che un tempo svolgeva questa funzione, era scomparso e la lince lo aveva sostituito nell'ecosistema semplice dell'isola.

Se la lince era in grado di regolare la popolazione dei caribù, si poteva ritenere che avesse lo stesso effetto anche sul numero delle lepri artiche. Nelle due zone di parto d'altopiano, dove maggiore era il tasso di sopravvivenza dei piccoli caribù, era stata segnalata sporadicamente anche la presenza di lepri artiche. D'altro canto era anche possibile che le lepri scarpa da neve avessero contribuito a ridurre la popolazione delle lepri artiche locali a causa di una concorrenza alimentare. Poteva anche essersi verificato un mutamento nella flora dell'isola che avrebbe eliminato la fonte alimentare delle lepri autoctone.

Per verificare se la moltiplicazione delle lepri artiche fosse stata condizionata dalla disponibilità di cibo due coppie di maschi e femmine sono state liberate sulla Brunette Island nella Fortune Bay circa 16 chilometri a sud di Terranova. Nell'isola non esistevano mammiferi predatori o lepri scarpa da neve. Inoltre, come buona parte dell'isola di Terranova, anche questa era costituita da zone alpine e subalpine e le sommità dei rilievi della tundra erano ricoperte da muschi e da ericacee. Dopo sei stagioni riproduttive, l'isola risultò popolata da 1000 lepri artiche.

Il ripopolamento della Brunette Island aveva dimostrato che il fattore cibo non costituiva una barriera per la crescita delle popolazioni di lepri artiche. Pertanto abbiamo deciso di utilizzare queste stesse lepri per verificare le altre ipotesi e le abbiamo introdotte in isole abitate da lepri scarpa da neve, ma non da linci. Come controllo sperimentale ne abbiamo liberate altre in zone di Terranova dove erano presenti le lepri scarpa da neve e le linci.

Nelle isole remote, dove erano presenti le lepri scarpa da neve ma non le linci, le lepri artiche sono sopravvissute e la loro popolazione è aumentata, anche se meno rapidamente di quanto si era verificato sulla Brunette Island, mentre negli habitat di Terranova, frequentati dalle linci, nessuno di questi ripopolamenti ha avuto successo. Era chiaro che, almeno nei bassopiani di Terranova, la lepre artica locale era in grado di sostenere la concorrenza della lepre scarpa da neve, ma non la predazione della lince.

Le verifiche sul campo a Terranova hanno dimostrato che la lince limitava l'abbondanza sia dei caribù sia delle lepri autoctone. A seguito degli esperimenti di ripopolamento e di uno studio effettuato sulle documentazioni storiche relative alla fauna dell'isola, hanno cominciato a delinearsi i contorni di un inconsueto sistema dinamico comprendente il caribù, la lince e le due specie di lepri. Per migliaia di anni prima dell'insediamento dell'uomo bianco a Terranova, le lepri artiche, i lupi e i caribù avevano coesistito sulle steppe ventose e sui massicci dell'isola. All'inizio la lince era poco comune, anzi, talmente rara che esperti osservatori non sono d'accordo se sia il caso di considerarla una specie indigena.

Indubbiamente, le popolazioni di lupi, caribù, linci e lepri artiche erano fluttuanti, ma i meccanismi stabilizzatori riuscivano a contenere queste fluttuazioni entro margini ristretti, così da evitare l'estinzione di una qualsiasi di queste specie. Successivamente, nel 1864, furono introdotte le lepri scarpa da neve per contribuire all'alimentazione dei gruppi isolati di pescatori della costa occidentale di Terranova. L'introduzione di questo nuovo tipo di lepre ruppe l'equilibrio della fragile dinamica di popolazione che si era imposta fra le specie indigene.

Quando un animale come la lepre scarpa da neve arriva in un nuovo ambiente, tende spesso a riprodursi rapidamente e, infatti, pochi decenni dopo la popolazione introdotta aveva raggiunto il suo massimo livello. Nel 1890, era stata raggiunta in tutta l'isola la punta massima.

Il picco iniziale conseguente all'introduzione di una specie è seguito generalmente da un declino, quando i limiti insiti della nicchia ecologica occupata cominciano a esercitare i propri effetti; così, verso il 1915, la popolazione di lepri scarpa da neve aveva iniziato un rapido calo.

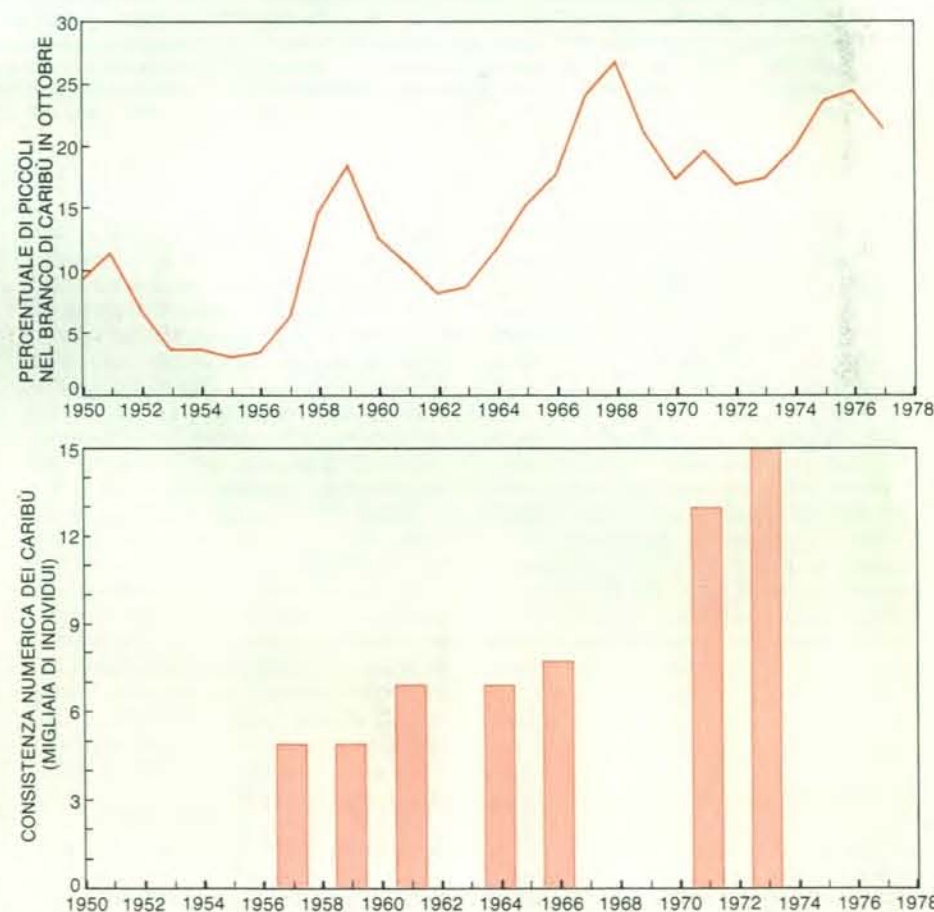
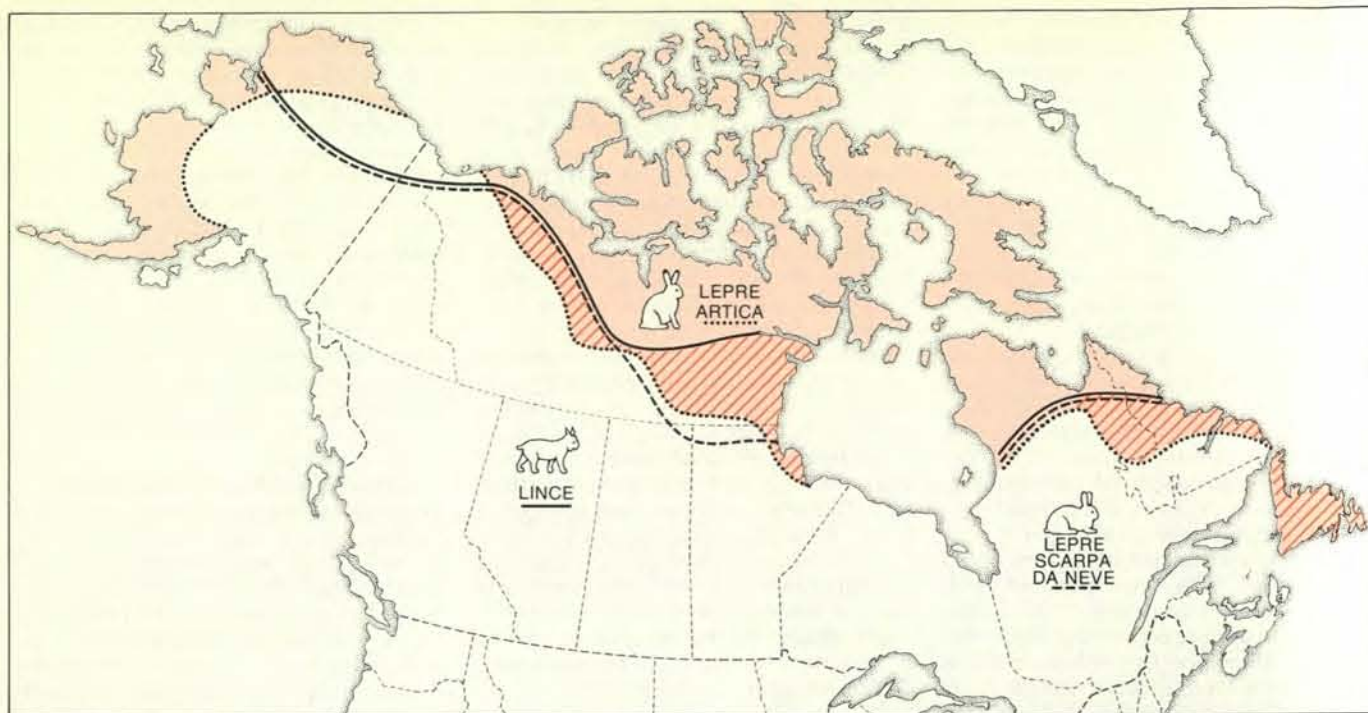
In gran parte del Canada, la lince è il principale predatore della lepre scarpa da neve. Quando, verso la fine dell'Ottocento, in quel paese la popolazione di quest'ultima specie andava aumentando con ritmo incessante, anche la popolazione delle linci seguiva la stessa tendenza, sebbene la sua curva d'incremento non mostrasse indici così impressionanti. È questo uno schema tipico nelle zone in cui la lince e la lepre scarpa da neve sono legate da un forzoso rapporto preda-predatore. In queste regioni, le curve che rappresentano le due popolazioni assumono una forma sinusoidale regolare. La popolazione di lepri scarpa da neve aumenta finché è limitata o dalla scarsità di cibo o da un meccanismo biologico intrinseco. In seguito diminuisce in quanto si abbassa il suo tasso riproduttivo. Quando le lepri diventano scarse, cala anche il tasso riproduttivo delle linci e quindi la loro popolazione. In assenza di interferenze esterne il ciclo completo comprende dai nove agli 11 anni.

Verso i primi anni del XX secolo la lince si era diffusa dalla costa occidentale su tutto il territorio di Terranova ed erano stati segnalati esemplari anche entro il perimetro della città di St. John, la capitale sita sulla penisola di Avalon, che si protende dalla costa sudorientale. Nel 1904 fu inoltrata al Parlamento di Terranova una proposta di legge intesa a sterminare questi animali, il che dimostra quanto rapidamente si fossero riprodotti. Alla fine la popolazione fu regolata non da fattori umani ma da fattori ecologici.

La popolazione di lepri scarpa da neve si stabilizzò sul valore massimo fra il 1896 e il 1915 (le oscillazioni delle popolazioni di lepri e di linci non si erano ancora consolidate), ma quando subì un crollo le linci furono private della loro principale risorsa alimentare. Questi animali sono però «spazzini» pieni di risorse e predatori molto flessibili e, quando la scarsità di cibo si fa sentire in modo marcato, interrompono la riproduzione ma raramente soffrono la fame. Si rivolgono, invece, ad altre prede: specie selvatiche o animali domestici. Oppure si nutrono dei rifiuti che dal mare si arenano sulle spiaggie.

A Terranova, le uniche prede selvatiche alternative erano topi, lepri artiche e piccoli di caribù. Forse le linci cominciarono a predare questi ultimi e le poche lepri artiche disponibili: le nuove prede non erano tanto abbondanti da consentire loro di aumentare o almeno di mantenere il loro livello numerico, ma di fatto servirono ad attenuare gli effetti della scomparsa della principale fonte alimentare.

Pressappoco nello stesso periodo in cui le linci cominciarono a catturare un gran numero di giovani caribù, il tasso di mortalità dei caribù adulti si elevò per tutt'altra ragione. La ferrovia di Terranova, completata verso il 1900, attraversa il sentiero che era sempre stato seguito dai branchi di caribù durante la loro migrazione autunnale in direzione sud. Dal 1911 al 1925 i cacciatori, servendosi della linea ferroviaria per intercettare gli animali, cominciarono a cacciare i capi in soprannumero e in ciascuno degli anni 1911, 1912, 1914 e 1915 abbatterono più di 5000 esemplari.

Dato che questa drastica decimazione dei caribù adulti si verificava nello stesso momento in cui le linci iniziavano a predarne i piccoli, non vi erano sufficienti nuove presenze per reintegrare le perdite provocate dai cacciatori. In pochi anni furono sterminati quasi tutti i 40 000 caribù presenti nel 1900. Nel corso dei due decenni in cui avvenne la catastrofica ri-

Negli anni cinquanta e sessanta il tasso di sopravvivenza dei giovani caribù ha determinato la consistenza del branco principale. La curva in alto si riferisce alla parte di piccoli che facevano parte del branco in ottobre e che avevano quindi superato la primavera e l'estate. I massimi di sopravvivenza rilevati nel 1951, 1959, 1968 e 1976 coincidono con i massimi di incremento numerico delle lepri scarpa da neve sull'isola. Quando il numero di queste ultime diminuiva, diminuiva anche il tasso di sopravvivenza dei piccoli caribù. Gli istogrammi in basso mostrano la consistenza totale del branco nella parte interna dell'isola di Terranova negli anni per i quali sono disponibili dati. Quando il ciclo di avvicendamento delle prede raggiungeva la massima intensità, l'entità del branco veniva determinata dalla sopravvivenza dei piccoli. Di recente l'intensità del ciclo si è attenuata per la diminuzione delle linci e per l'introduzione di nuove specie di prede.

L'area di distribuzione della lepre artica in Canada (*in colore*) potrebbe essere limitata in parte dai rapporti preda-predatore. Il suo limite meridionale coincide con il limite settentrionale dell'area di distribuzione della lince e le due aree si sovrappongono solo parzialmente (*zona tratteggiata*). Al contrario, le aree di distribuzione della lince e della lepre scarpa da neve si sovrappongono in misura notevole. Questi due animali hanno poco carico sulle zampe, ossia la pressione che esercitano sul terreno quando corrono è scarsa. Essi sono quindi ben adattati alla corsa sulla neve soffice della foresta boreale. Abituata alla corsa sulla neve gelata della tundra, la lepre artica ha un maggior carico e nella foresta procede a balzi, esponendosi all'attacco della lince. In assenza di linci le lepri artiche potrebbero spingersi molto più a sud.

duzione dei loro branchi la popolazione di lepri scarpa da neve aveva cominciato a fluttuare secondo lo schema ciclico osservato in altre regioni del Canada. Furono riscontrate significative punte negli anni 1920, 1931, 1940-1943, 1951-1952, 1959-1960, 1969 e 1976.

Parallelamente ai picchi del ciclo delle lepri scarpa da neve, le linci ripresero a predarle con conseguente incremento dei tassi di sopravvivenza fra i giovani caribù. Le statistiche dimostrano che i branchi di caribù aumentarono rapidamente attorno al 1940, 1950 e 1960. In corrispondenza del picco del 1960 per la lepre scarpa da neve, i principali branchi di caribù che sono stati conteggiati sono aumentati dai 4800 capi del 1959 ai 6100 del 1961. L'adattamento predatorio della lince stava così a significare che le popolazioni di caribù e di lepri artiche erano strettamente correlate con la popolazione di lepri scarpa da neve.

Negli ultimi decenni, i cicli di sopravvivenza dei giovani caribù sono diventati meno accentuati e anche durante le fasi calanti del ciclo è sopravvissuta, rispetto alle fasi più intense, una quota maggiore di giovani individui. Quest'attenuazione si è verificata in parte perché la popolazione di linci sull'isola è diminuita a causa di un altro fenomeno ciclico: la moda femminile.

Dopo un certo intervallo di tempo, le pellicce a pelo lungo sono ritornate di moda, rendendo economicamente vantaggiosa la cattura delle linci. La densità della popolazione di linci a Terranova è così diminuita. Inoltre, sono state introdotte sull'isola nuove specie di prede, fra cui i tetraonidi *Bonasa umbellus* e *Canachites canadensis*. La catena alimentare dell'isola è divenuta allora più complessa, aggiungendo un effetto mitigatore alle oscillazioni delle popolazioni predatrici.

Gli effetti negativi prodotti dalla lince sui caribù di Terranova derivano principalmente dal fatto, che per un lungo periodo di tempo, i caribù si sono adattati alla predazione da parte del lupo. Le strategie dimostratesi efficaci contro questo animale diventavano però un inconveniente nel caso della comparsa di un nuovo predatore che avesse abitudini di caccia molto diverse. Per esempio, i grossi branchi di caribù che si formano a primavera dopo il periodo delle nascite forniscono un vantaggio sostanziale nella difesa dei piccoli contro il lupo, che caccia spesso di giorno e insegue la preda in spazi aperti. Quando un branco di caribù vede un lupo, la sua principale strategia consiste nel cercare scampo nella fuga. L'occhio del caribù è, inoltre, ben adattato a percepire i movimenti durante il giorno, un prezioso vantaggio nei riguardi di un predatore diurno che si sposti all'aperto. Per la stessa ragione, lo spazio libero dove le femmine partoriscono favorisce l'individuazione dei lupi che vanno a caccia di prede.

Questa combinazione di aspetti fisiologici e comportamentali si dimostra però di scarsa utilità quando il predatore è la lince, che caccia prevalentemente di notte e attacca la preda all'agguato invece di catturarla in piena corsa. Le linci usano acquattarsi e restare immobili quando sono alla posta e questo loro atteggiamento neutralizza le difese del caribù, i cui occhi sono adattati a individuare più il movimento che la forma. Generalmente il caribù non ha reazioni di allarme in presenza di un animale immobile.

Quando un predatore insegue un gruppo di erbivori, sceglie spesso di attaccare un particolare animale che si differenzia in qualche modo dagli altri. I lupi scelgono caribù giovani, malati, vecchi o azzoppati e praticamente qualsiasi differenza di spicco può attirare la loro attenzione. Invece la selezione delle linci è concentrata esclusivamente sui piccoli; l'obiettivo più attraente per questi animali è un giovane esemplare vivace e curioso, che si avvicini ai margini della radura dove è avvenuta la riproduzione delle femmine, presso la frangia di vegetazione che serve da riparo al predatore in azione.

Può darsi che la predazione da parte delle linci, in assenza dei lupi, abbia impresso una nuova direzione all'evoluzione dei caribù di Terranova. La selezione naturale può di solito favorire la dispersione dei piccoli più che la loro concentrazione nelle zone di parto, facendone comodi bersagli. Il legame fra madre e figlio si può, inoltre, rafforzare: il piccolo che non si allontana dalla madre ha più probabilità di sopravvivere alla predazione della lince di un altro che può correre velocemente, anche se quest'ultimo è favorito quando si tratta di sfuggire a un lupo.

Se il caribù si adatta, il suo tasso di sopravvivenza potrebbe accrescersi rapidamente, dato che il controadattamento della lince deve operare entro limiti più circoscritti. Dal punto di vista alimentare la lince non può contare unicamente su di lui, ma si rivolge prevalentemente alla lepre scarpa da neve. Qualsiasi modifica del suo comportamento di caccia non può alterare il quadro che ne ha fatto un'efficiente cacciatrice di lepri. Pertanto, un nuovo rapporto evolutivo fra lince e caribù dovrebbe favorire quest'ultimo.

Allo stesso modo del rapporto di predazione fra lince e caribù, il rapporto fra lince e lepre artica non si è affermato in seguito a un lungo periodo di stretta coevoluzione. Nella maggioranza dei casi, le due specie frequentano zone separate. Su una carta del Canada, il limite meridionale dell'area di distribuzione della lepre artica coincide puntualmente con il limite settentrionale dell'area di distribuzione della lince. Al contrario, i territori della lince e della lepre scarpa da neve si sovrappongono perlopiù ai margini: il predatore si spinge a nord come la preda.

La maggior parte degli ecologi spiegherebbe la distribuzione delle tre specie, ricorrendo al concetto di biomi o zone vitali: la lepre artica vive nel bioma della tundra, mentre la lince e la lepre scarpa da neve vivono nel bioma della foresta boreale. Anche i rapporti predatore-preda, però, potrebbero influenzare il campo di diffusione della lepre artica; nel Canada, questa specie, in assenza di linci, potrebbe sconfinare alquanto più a sud della sua abituale area di distribuzione.

Il successo della lince nel perseguire le due specie di lepri dipende, fra le altre cose, dalla vegetazione e dallo stato della neve. Negli habitat forestali la lince insegue e tende l'agguato alla sua preda mentre la lepre scarpa da neve si accuccia immobile e rimane in allerta. Il risultato di un incontro tra questi due animali dipende principalmente dalla distanza tra i due, quando la preda individua il predatore riuscendo a fuggire.

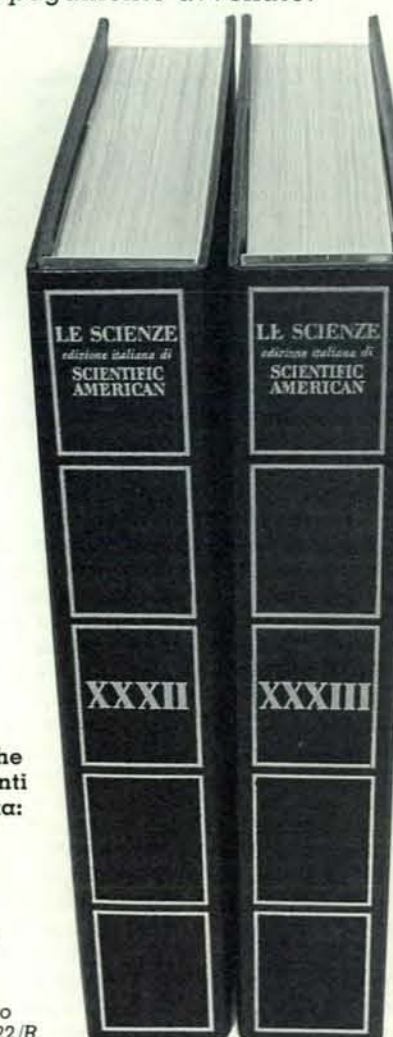
Mercer ha fatto notare che la lince e la lepre scarpa da neve sono in condizioni di parità sulla neve soffice della foresta boreale in quanto entrambe hanno poco carico sulle zampe, ossia esercitano una scarsa pressione sul suolo quando corrono. Quindi possono spostarsi rapidamente fra gli alberi senza correre il rischio di affondare nella neve.

Al contrario, la lepre artica è adattata alla neve gelata della tundra settentrionale e ha un carico più che doppio rispetto alla lepre scarpa da neve. Nella neve soffice procede a balzi irregolari, riducendo le sue probabilità di scampo rispetto alla lince. Questo non implica che sia unicamente lo stato della neve a determinare l'area di distribuzione della lepre artica. La combinazione fatale è costituita dalla presenza di neve soffice e di predatori con scarso carico sulle zampe.

I raccoglitori per il 1984

Questi raccoglitori corrispondono ai volumi XXXII e XXXIII de LE SCIENZE, e rispettivamente ai fascicoli da gennaio (n. 185) a giugno (n. 190) e da luglio (n. 191) a dicembre (n. 196).

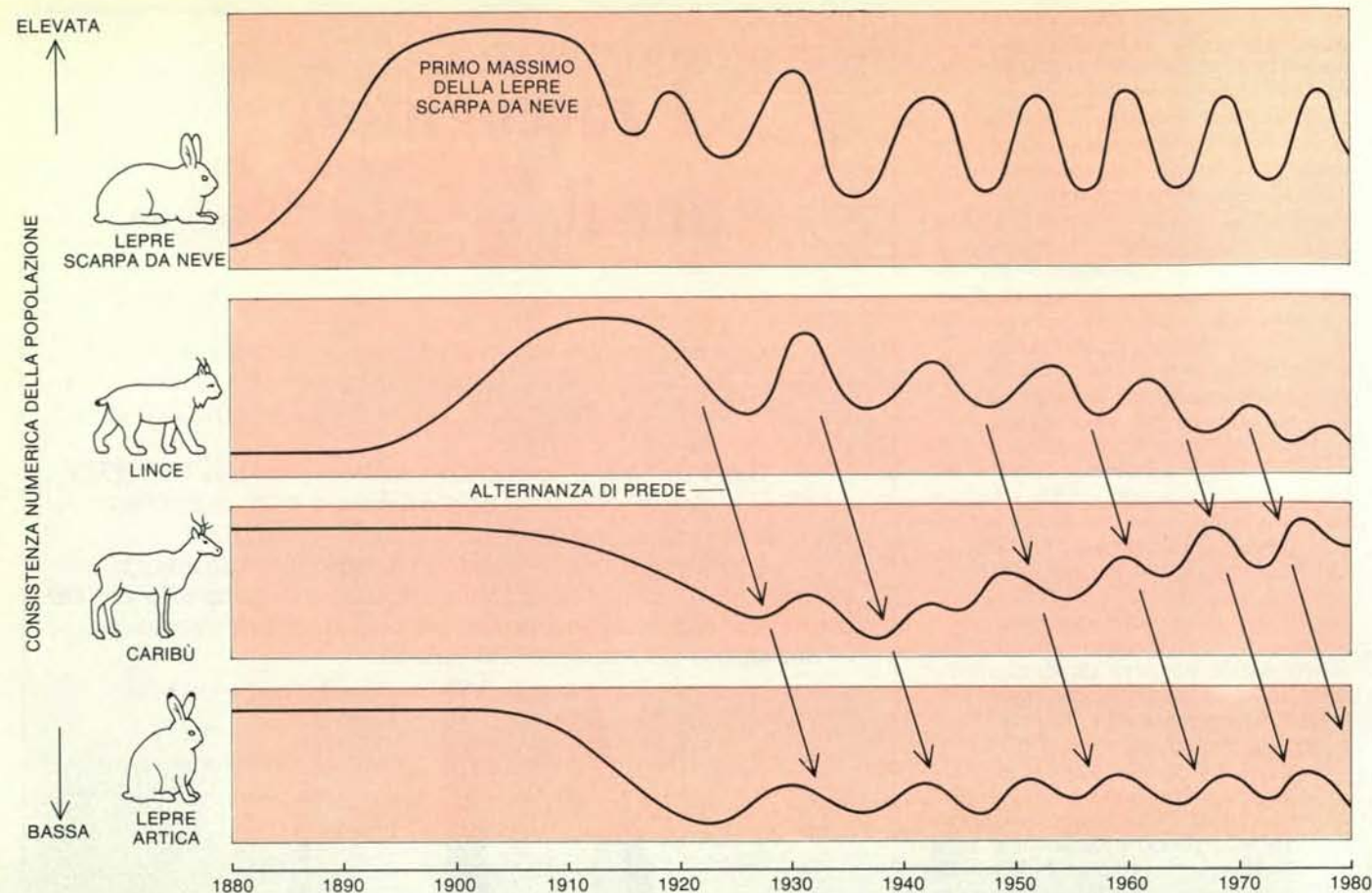
Sono ancora disponibili i raccoglitori dal Vol. XXIV al XXXI e dei raccoglitori non numerati appositamente approntati per sostituire i raccoglitori esauriti.

I raccoglitori si possono richiedere direttamente all'editore usando l'apposita cartolina allegata alla rivista e unendo il relativo importo; gli ordini infatti vengono evasi solo a pagamento avvenuto.

I raccoglitori si trovano anche presso i seguenti punti di vendita:

BOLOGNA
Libreria Parolini
Via U. Bassi 14
FIRENZE
Libreria Marzocco
Via de' Martelli 22/R
GENOVA
Libreria Int. Di Stefano
Via R. Ceccardi 40/R
MILANO
Le Scienze S.p.A.
Via del Lauro 14

TORINO
Libreria Zanaboni
C.so Vittorio Emanuele 41
NAPOLI
Libreria Guida A.
Via Port'Alba 20/21

PADOVA
Libreria Cortina
Via F. Marzolo 4
PALERMO
Libreria Dante
Quattro Canti di Città
ROMA
Claudio Aranci
Viale Europa 319 (EUR)

Ogni raccoglitore L. 3.600

La dinamica di popolazione della lepre scarpa da neve, della lince, del caribù e della lepre artica indicano che la forza motrice del sistema di alternanza di prede è il ciclo decennale proprio della lepre scarpa da neve. Il primo massimo numerico della popolazione di quest'ultima lepre a Terranova è stato il più alto; i successivi sono stati un poco più bassi. Prima dell'arrivo sull'isola delle lepri scarpa da neve, le linci erano talmente rare che si dubita ancora che questi animali siano una specie indigena. A ogni aumento della popolazione di lepri scarpa da neve ha sempre corrisposto un rapido incremento delle linci. Nel 1900 erano presenti sull'isola circa 40 000 caribù; nel 1925, a seguito della predazione delle linci e della caccia praticata dall'uomo, essi erano probabilmente ridotti a 1000 esemplari o meno. La popolazione di lepri artiche è nella «fossa del predatore»: continua a esistere perché è scarsa e non può uscire da questa situazione. Se essa aumentasse numericamente, le linci in breve tempo la ridurrebbero, non appena avesse inizio per le lepri scarpa da neve la fase declinante del ciclo.

L'avvicendamento dei predatori da una preda all'altra costituisce oggi un promettente campo d'indagine in ecologia. Questi eventi aiutano a spiegare in che modo si può conservare la stabilità in ecosistemi fragili e limitati. Quando scarseggia la preda primaria, un predatore efficiente può rivolgersi a una preda secondaria e, con questo meccanismo, mantiene la propria posizione in una rete alimentare che altrimenti crollerebbe. L'avvicendamento allevia anche la pressione predatoria su una popolazione predata in declino, permettendogli di sopravvivere. Ma anche così può dar luogo a una sopravvivenza della popolazione predata a bassissima densità. È quanto accade per la lepre artica a Terranova, dove l'alta densità delle linci, assicurata dalle lepri scarpa da neve, mantiene a basso livello la densità della lepre artica.

In un sistema di questo tipo, si usa dire che la lepre artica si trova nella «fossa del predatore». Un altro esempio di specie predata che subisce una situazione analoga è rappresentato dal caribù dell'Ontario e della British Columbia, regioni in cui questo animale costituisce un lato di un triangolo che comprende anche il lupo e l'alce americano. In entrambe il caribù si è evoluto parallelamente al lupo in assenza dell'alce. Nel corso del XIX secolo, questo si è diffuso, proveniendo da sud e da est. L'aggiunta di questo nuovo elemento alla disponibilità alimentare ha provocato un aumento della popolazione dei lupi. In confronto all'alce il caribù è però una preda più facile per il lupo che lo attacca appena può.

In alcune località dell'Ontario e della British Columbia i caribù sono, per usare la stessa espressione di sopra, «nella fossa del predatore», e si conservano solo perché sono rari. Se dovessero aumentare, il lupo ritornerebbe a predarli e il loro numero diminuirebbe. Non possono uscire da una simile situazione finché la popolazione di alci continua a sostenere l'alta densità dei lupi.

In alcuni casi, le specie predate in un simile sistema sopravvivono non solo perché sono poco numerose, ma anche perché la loro popolazione è talmente dispersa che cacciare i singoli animali risulta poco redditizio in termini di consumo energetico. Per esempio, l'attuale capacità di Terranova di sostenere la lepre artica è definita dall'area minima per lepre che è indispensabile a rarefare le possibilità d'incontro fra le linci e questa preda.

Questi incontri devono essere, di fatto, sufficientemente sporadici perché la sopravvivenza delle giovani lepri artiche compensi la mortalità fra gli adulti. Se aumentasse il loro numero, e quindi la loro densità, la situazione sarebbe compromessa e la predazione da parte delle linci ristabilirebbe presto l'equilibrio. Stando così le cose, le relazioni spaziali si rivelerebbero come aspetto altamente significativo della dinamica dei sistemi predatore-preda.

La conclusione più interessante che emerge dallo studio della fauna di Terranova è, però, il considerevole e imprevedibile effetto che può essere causato dall'introduzione di una nuova specie in un ecosistema semplice. Quanti fra gli abitanti del 1864 avrebbero potuto prevedere che l'importazione di poche lepri scarpa da neve, che dovevano servire a nutrire un gruppo di pescatori affamati, a distanza di 100 anni avrebbe condizionato ciclicamente la sopravvivenza dei giovani caribù e limitato la popolazione delle lepri artiche?

Le inclusioni fluide nei minerali

Lo studio condotto su queste «gocce» intrappolate da milioni di anni nei minerali fornisce informazioni di interesse teorico e pratico sui giacimenti minerari, sui campi geotermici e sui processi magmatici

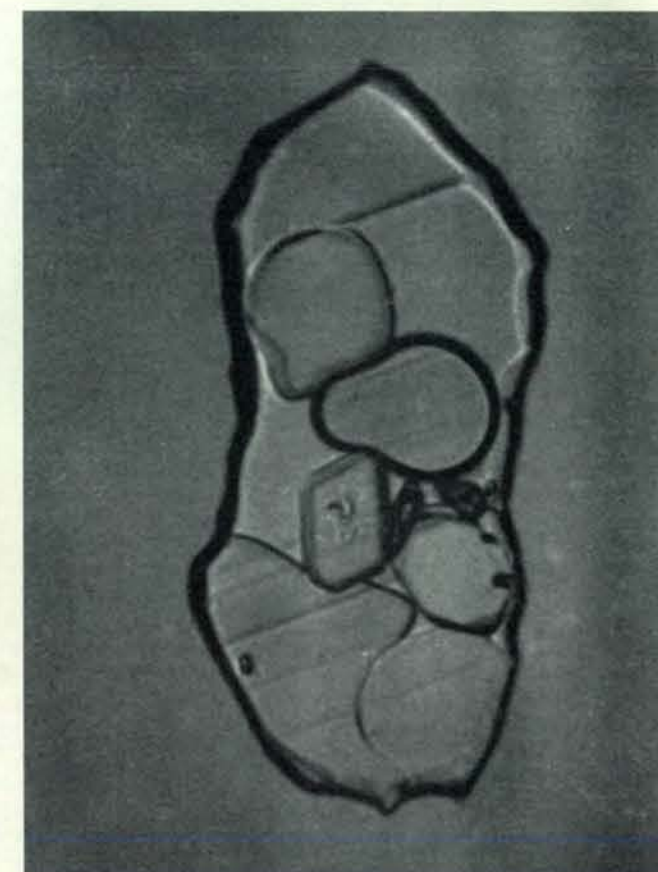
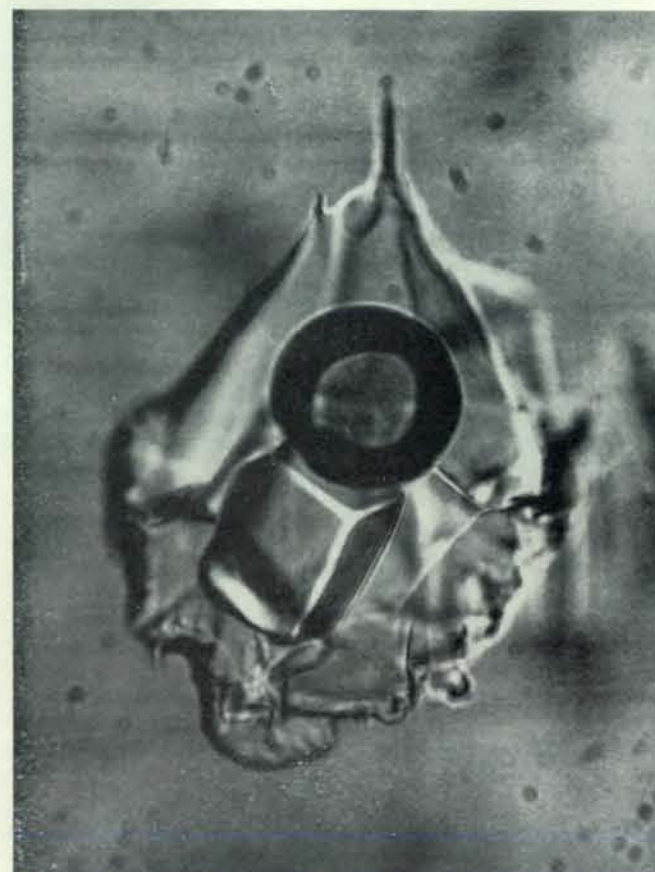
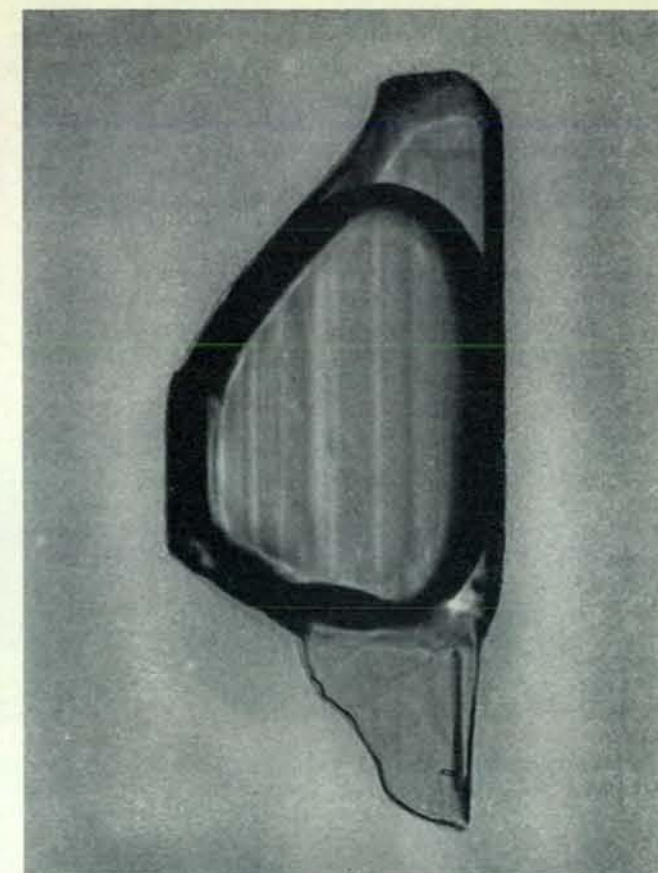
di Benedetto De Vivo

È noto che gran parte delle rocce sedimentarie si sono formate in acque marine; non è invece altrettanto noto che molti processi geologici che hanno luogo nell'ambito della crosta terrestre, ivi compresa la formazione di molti giacimenti minerari, avvengono in ambienti permeati da acque saline. La chiave di lettura di molti di questi processi è rappresentata da goccioline di antiche soluzioni, le quali si sono conservate entro molti minerali e rocce, di diversa provenienza geologica, sotto forma di «inclusioni fluide», ossia gocce di soluzioni acquose, gassose o di altri fluidi (inclusi i fusi silicatici). Le inclusioni sono vere e proprie «bottiglie» in cui sono contenuti fluidi rimasti intrappolati anche per milioni di anni e che possono dare informazioni riguardanti la temperatura, la pressione, la densità e la composizione dei fluidi esistenti in quel determinato ambiente geologico in epoche geologiche passate. Le inclusioni hanno dimensioni variabili, comprese tra un micrometro (o anche meno) e alcuni centimetri. Quelle con diametro maggiore di un millimetro sono abbastanza rare; quelle nell'ordine dei centimetri sono «pezzi da museo». Le inclusioni fluide sono facilmente osservabili, al microscopio, in minerali bianchi e traslucidi, quali quarzo, calcite, fluorite; spesso, in minerali di questo tipo, si può avere una concentrazione di milioni di inclusioni per centimetro cubo. Spesso le inclusioni, osservate con un microscopio a luce trasmessa, appaiono come sferule completamente nere a causa della riflessione totale della luce da parte delle pareti delle inclusioni stesse. Quando invece hanno una forma appiattita riescono a trasmettere molto bene la luce. Molte inclusioni contengono soluzioni con concentrazioni saline variabili fra quelle di acque dolci e quelle di acque ad altissima salinità. Alcune sono composte da una fase liquida, una fase vapore e una fase solida; altre possono essere costituite da due liquidi immiscibili; altre ancora possono essere formate da una sola fase gassosa. La grande abbondanza e la diffusione delle inclusioni fluide stanno a indicare che molti processi geologici hanno avuto luogo in ambienti in cui microfratture e pori erano pieni di liquidi. Rocce prive di inclusioni, quali le meteoriti e alcuni tipi di rocce ignee e metamorfiche, si sono formate in condizioni speciali caratterizzate dalla mancanza di fluidi.

Le inclusioni fluide sono state riconosciute e studiate da più di 100 anni, ma già Robert Boyle nel 1672 segnalava l'esistenza entro un cristallo di quarzo di grosse bollicine piene di un liquido in movimento. I primi lavori analitici che hanno cercato di stabilire la composizione delle inclusioni si devono a diversi ricercatori dell'inizio del XIX secolo quali Scipione Breislak, Sir Humphry Davy, il fisico scozzese Sir David Brewster e William Nichol anch'egli scozzese. Si devono in particolare all'inglese Henry Clifton Sorby molte osservazioni, tuttora valide, sulla natura di tali inclusioni compiute verso la metà del secolo scorso. Tra l'altro Sorby stabiliva che le bolle presenti nei fluidi di molte inclusioni rappresentano il prodotto di una contrazione differenziale del liquido e del cristallo ospite verificatasi durante il raffreddamento dalla temperatura di intrappolamento dell'inclusione alla temperatura di osservazione. Negli anni successivi molte delle intuizioni di Sorby e di un altro suo eminente contemporaneo, il petrografo tedesco Ferdinand Zirkel, furono riprese e utilizzate da più ricercatori per uno studio in senso moderno delle inclusioni e per una loro utilizzazione in diversi problemi di carattere applicativo.

In tempi più recenti la tecnica delle inclusioni fluide è stata particolarmente utilizzata da ricercatori della scuola sovietica specialmente in problemi connessi con lo studio della genesi di giacimenti minerari. N. P. Ermakov dell'Università statale di Mosca è stato senz'altro colui che in Unione Sovietica ha dato maggiore impulso a questo tipo di studi e va considerato il caposcuola nei paesi dell'Est europeo. Nei paesi occidentali, invece, il più grosso contributo allo studio delle inclusioni fluide si deve al geologo Edwin Roedder dell'US Geological Survey, a cui va ascritto il merito di aver diffuso e «propagandato» anche fra i più scettici tale metodologia con una serie di pregevoli lavori. Un contributo rilevante allo studio delle inclusioni fluide è venuto infine anche da parte della scuola francese e tra i tanti ricercatori è doveroso ricordare il ruolo svolto dal mineralogista Georges Deicha e da Bernard Poty del Centre de Recherche sur la Geologie de l'Uranium di Vandoeuvre-lès-Nancy in Francia.

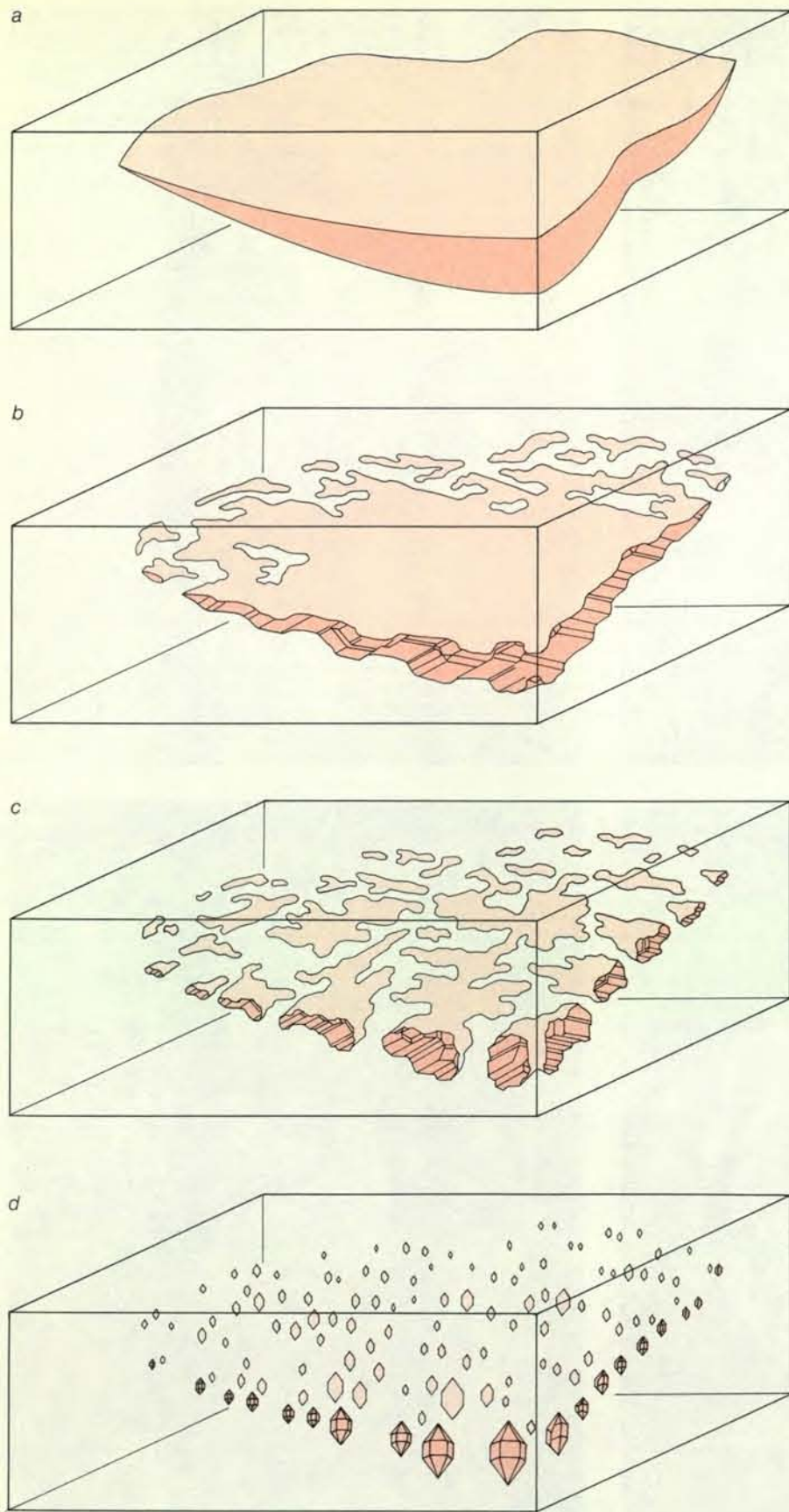
A seconda della loro origine le inclusione fluide vengono classificate in primarie, secondarie e pseudosecondarie. Le prime si formano durante la crescita del minerale ospite, per la presenza di irregolarità e di disomogeneità nel fluido e per una qualsiasi altra causa che interferisca con la crescita del cristallo. Le inclusioni secondarie si formano in qualche momento successivo all'avvenuta cristallizzazione del minerale ospite in relazione a imprecisate tensioni a cui le rocce della crosta vengono sottoposte. In questo caso, se un cristallo si frattura in presenza di un fluido, quest'ultimo può penetrare nella frattura avviando un processo di soluzione e ricristallizzazione del cristallo ospite che porterà a una riduzione della superficie libera a disposizione e alla formazione di una serie di inclusioni man mano che il processo progredisce.

Le inclusioni pseudosecondarie rappresentano un caso intermedio rispetto a quanto descritto per le inclusioni primarie e per quelle secondarie. Esse infatti si formano quando un cristallo si frattura mentre è ancora in fase di crescita, permettendo così ai fluidi di penetrare nella frattura e di restarvi intrappolati nel momento in cui avviene la ricristallizzazione. Le inclusioni pseudosecondarie sono apparentemente del tutto simili alle

In una inclusione fluida il numero di fasi varia. L'area oblunga scura nella microfotografia in alto a sinistra è mercurio metallico liquido in calcite. L'inclusione in alto a destra in acquamarina è invece a due fasi: una bolla di gas e sopra di essa un volume più piccolo di liquido. L'inclusione a tre fasi, in basso a sinistra, è in smeraldo e consiste di un cristallo cubico di salgemma e di una bolla di anidride carbonica circondati da acqua gelata. L'inclusione in basso a destra è a più fasi: un cristallo di magnesite, una bolla di gas, una fase liquida e sei fasi solide. L'ingrandimento delle microfotografie, di Edwin Roedder dell'US Geological Survey, è rispettivamente di 27, 277, 520 e 366 diametri.

a

b

c

d

A differenza delle inclusioni fluide primarie che si formano contemporaneamente alla crescita del minerale ospite, le inclusioni secondarie si formano in un secondo tempo entro la frattura di un cristallo (*a*) che si trovi immerso in un liquido (*in colore intenso*). La soluzione e la rideposizione di materiale dal liquido sulle superfici della frattura danno luogo a forme di accrescimento di tipo dendritico (*b*) che inglobano piccoli volumi di liquido (*c*); la loro area superficiale in seguito diminuisce ed essi assumono la forma di masse rotondeggianti o di cristalli negativi cavi (*d*).

inclusioni secondarie; tuttavia dal punto di vista della genesi, e quindi delle informazioni che da esse derivano, devono essere considerate come vere e proprie inclusioni primarie. Una chiara distinzione fra inclusioni secondarie e inclusioni pseudosecondarie, però, risulta spesso molto difficile, se non addirittura impossibile.

Nei paragrafi che seguono, dopo aver descritto quali sono i parametri derivabili dalle inclusioni fluide, ossia temperatura, pressione, densità e composizione chimica, illustrerò alcune applicazioni pratiche a problemi connessi con le scienze della Terra.

Sperimentalmente si determina di solito la temperatura di omogeneizzazione, la quale è correlata alla temperatura di formazione o di intrappolamento dell'inclusione, definita anche temperatura di omogeneizzazione con correzione di pressione. Quando, dopo la formazione di una inclusione entro un cristallo la temperatura si abbassa, il fluido caldo dell'inclusione si condensa da una fase supercritica a una fase in cui coesistono liquido e vapore. In laboratorio, per mezzo di un tavolino riscaldante applicato a un normale microscopio da mineralogia, si inverte questo processo, riscaldando di nuovo l'inclusione, e si osserva che a una certa temperatura la fase vapore (oppure liquida) scompare. La temperatura alla quale ha luogo questo fenomeno si definisce temperatura di omogeneizzazione. Se la pressione esterna che agiva sul cristallo contenente l'inclusione al momento in cui l'inclusione è stata intrappolata era maggiore della pressione di vapore che caratterizza il fluido contenuto nell'inclusione nel momento in cui in laboratorio si ottiene la temperatura di omogeneizzazione, l'omogeneizzazione dell'inclusione avrà luogo a una temperatura inferiore a quella a cui l'inclusione si era inizialmente formata, ossia alla cosiddetta temperatura di formazione. La differenza che bisogna aggiungere alla temperatura di omogeneizzazione per ottenere la temperatura di formazione è la «correzione di pressione». Tali correzioni si ricavano dalle tabelle sperimentalmente ottenute da Robert W. Potter dell'US Geological Survey. Tale correzione può essere zero, se l'inclusione si è formata da un fluido che è rimasto intrappolato allorché si trovava in fase di ebollizione (quindi la pressione che agiva sul cristallo contenente l'inclusione al momento dell'intrappolamento era uguale alla pressione di vapore del fluido nell'inclusione), oppure è dell'ordine di poche decine di gradi se si tratta di fluidi epitermali, ossia di bassa temperatura. Se invece fluidi a densità relativamente bassa (0,4 - 0,6 grammi per centimetro cubo) vengono intrappolati in regime di alta pressione (nell'ordine cioè di 2-5 chilobar), allora la correzione di pressione può raggiungere anche centinaia di gradi.

La temperatura di omogeneizzazione delle inclusioni fluide rappresenta solo un valore minimo della temperatura di formazione, ma poiché la prima è funzione sia della temperatura sia della pressione di formazione, è necessario un metodo indipendente per ottenere la pressione.

Generalmente sulla base di ricostruzioni e di dati geologici si effettua una stima della profondità a cui è avvenuta la formazione delle inclusioni, stima che, accoppiata ai dati sulla composizione del fluido contenuto nell'inclusione, permette di determinare la correzione di pressione da apportare alla temperatura di omogeneizzazione per ottenere la temperatura di formazione applicando le tabelle di Potter. Oltre che a queste due temperature si fa spesso riferimento, soprattutto nella letteratura sovietica, anche alla «temperatura di decrepitazione» assunta molto spesso arbitrariamente come equivalente della temperatura di formazione. La temperatura di decrepitazione viene misurata per mezzo di decrepitometri. Il metodo consiste nel riscaldare abbastanza velocemente minerali (o rocce) preventivamente frantumati e nel raccogliere con appositi microfoni gli impercettibili suoni emessi per effetto della decrepitazione dei grani; i suoni poi vengono amplificati e riportati in diagramma sotto forma di intensità in rapporto alla temperatura. La decrepitazione può avvenire per una serie di fattori non facilmente controllabili, ma si assume che fondamentalmente sia dovuta al fatto che la pressione interna esistente nelle inclusioni fluide a una certa temperatura superi la resistenza meccanica del minerale ospite e lo rompa. Come si può intuitivamente capire non è affatto detto che la temperatura di decrepitazione coincida con la temperatura di formazione, data la diversità dei fattori che contribuiscono a determinare la rottura del cristallo. Ne deriva quindi che il metodo decrepitometrico non è completamente affidabile come «geotermometro»; oltretutto è un metodo indiretto e quindi non comparabile al metodo microscopico dell'osservazione diretta dell'inclusione attraverso il quale si ottiene la temperatura di omogeneizzazione. Il metodo decrepitometrico ha tuttavia la sua validità come strumento di campagna, nelle fasi esplorative per fare rapide e grossolane distinzioni sul terreno di aloni di inclusioni di alta temperatura, ai quali possono essere eventualmente associati depositi minerari.

In letteratura sono descritti diversi metodi che, utilizzando solo i dati derivanti dallo studio delle inclusioni fluide, permettono di stimare la pressione. Alcuni di questi metodi, comunque, forniscono soltanto informazioni sulla pressione minima necessaria per mantenere un determinato fluido alla densità determinata sperimentalmente alla temperatura di intrappolamento. Questo avviene quando si è in presenza, per esempio, di un fluido omogeneo con inclusioni liquido-vapore aventi la stessa composizione e stessa temperatura di omogeneizzazione. In questo caso si può assumere che la pressione idrostatica, al momento dell'intrappolamento delle inclusioni, fosse più elevata della pressione di vapore di quella particolare soluzione a quella temperatura, altrimenti il fluido sarebbe entrato in ebollizione. In questo caso, quindi, si può fissare solo la pressione minima che doveva agire su quel determinato fluido affinché non si avesse ebollizione. Va detto che questo tipo di inclusioni (liquido + vapore) rappresenta pressoché la norma nella maggior parte dei campioni. Una volta note la composizione e la temperatura di omogeneizzazione è possibile ottenere da queste inclusioni anche informazioni sulla pressione assoluta più dettagliate di quelle ottenibili con normali metodi petrografici, se si dispone di un geotermometro indipendente che dia indicazioni sulla temperatura di formazione dei minerali ospiti e se si fa uso dei diagrammi sperimentali pressione-volume-temperatura relativi a fluidi di composizione nota. Altri metodi, invece, forniscono informazioni sulla pressione assoluta esistente al momento della formazione delle inclusioni, utilizzando, per esempio, le inclusioni in cui è rimasto intrappolato un liquido in «ebollizione» e la fase vapore coesistente. Anche nelle inclusioni che intrappolano separatamente la fase liquida e la fase vapore, si formano poi, per raffreddamento, due fasi: nell'inclusione contenente la fase liquida, infatti, apparirà, per contrazione, una bolla di vapore, mentre nell'inclusione contenente la fase vapore si svilupperà, per condensazione, una pellicola di liquido. Riscaldando questi due tipi di inclusione, l'omogeneizzazione, nella fase liquida e nella fase vapore, rispettivamente, avviene necessariamente alla stessa temperatura e, se è nota la curva di ebollizione del particolare fluido contenuto nell'inclusione, si può risalire dalla temperatura di omogeneizzazione alla pressione che esisteva al momento della formazione dell'inclusione. Per poter utilizzare il metodo descritto è fondamentale determinare se coesistono inclusioni con una fase vapore predominante e inclusioni con una fase liquida predominante a testimonianza che il fluido al momento dell'intrappolamento nell'inclusione si trovava in stato di ebollizione. L'individuazione dello stato di ebollizione ha un interesse non puramente speculativo, in quanto è noto che molti adunamenti minerari si formano in seguito a importanti variazioni fisico-chimiche che subiscono le soluzioni mineralizzanti in conseguenza appunto dell'ebollizione.

L'applicazione del metodo descritto è valida soprattutto per il sistema acqua-cloruro di sodio; per quanto riguarda invece la determinazione della pressione sulla base di inclusioni contenenti anidride carbonica rimandiamo la trattazione a più avanti, quando parleremo dei noduli pirossenitici del Vesuvio.

Lo studio delle inclusioni fluide costituisce un metodo accurato per la determinazione delle densità dei fluidi presenti durante processi geologici verificatisi milioni di anni addietro. Una volta che sono state determinate le fasi presenti nelle inclusioni in esame e le relative temperature di omogeneizzazione, è facile, attraverso diagrammi sperimentali, ottenere dati

È possibile derivare la temperatura di intrappolamento, o di formazione, di una inclusione fluida dalla temperatura di omogeneizzazione, dalla composizione della inclusione e dalla pressione esistente al momento della formazione. Anche se comunemente le inclusioni fluide contengono sali diversi, è buona approssimazione esprimerne la salinità in percentuale di cloruro di sodio equivalente. Il grafico si riferisce a soluzioni con salinità del 5 per cento di cloruro di sodio equivalente; i valori sulle curve danno la pressione in megapascal, mentre sull'ordinata è riportata la lettura della correzione di temperatura da aggiungere (o sottrarre) alla temperatura di omogeneizzazione per ottenere la temperatura di formazione. Il grafico è basato su dati sperimentali ottenuti da Robert W. Potter dell'US Geological Survey.

sulla densità dei fluidi di cui si conosce la composizione. I dati sulla densità sono anche fondamentali per i metodi adottati per stimare la pressione esistente sulle inclusioni al momento della loro formazione e assumono una importanza notevole nello studio dei fluidi mineralizzanti, in quanto la densità costituisce uno dei parametri fondamentali che controllano il movimento dei fluidi.

Il liquido che normalmente è contenuto nelle inclusioni è una soluzione acquosa, con concentrazioni che possono variare tra il 50 per cento in peso di sali e praticamente lo zero per cento. I sali sono composti principalmente da ioni sodio, potassio, calcio, magnesio, cloro, solfato con quantità subordinate di ioni litio, alluminio, borato, fosfato, metasilicato acido, bicarbonato, carbonato e altri ioni di minore importanza. Gli ioni che nella generalità predominano sono comunque quelli sodio e cloro. La presenza di anidride carbonica non è infrequente; essa talvolta anzi rappresenta, nelle inclusioni, il gas predominante.

Come abbiamo visto, la bolla che normalmente si trova nelle inclusioni si forma in conseguenza alla contrazione differenziale del liquido e del cristallo ospite durante il raffreddamento. Ne deriva che la bolla può essere costituita o da vapore acqueo o da gas che si trovavano originariamente disciolti nel fluido intrappolato. Quando il gas è prevalentemente anidride carbonica si formano, nell'inclusione, per raffreddamento, due liquidi immiscibili (soluzione acquosa e anidride carbonica liquida) e una bolla di gas (anidride carbonica sotto pressione). Più raramente può anche succedere che si osservino due liquidi immiscibili di cui l'uno è costituito da idrocarburi e l'altro da una soluzione acquosa; una tale inclusione deriva dall'intrappolamento di una emulsione di idrocarburi fluidi nel fluido acquoso.

Se, durante il raffreddamento, il fluido contenuto nell'inclusione diventa saturo rispetto a uno dei sali disciolti, come per esempio cloruro di sodio, possono formarsi nel fluido nuovi cristalli chiamati «minerali figli» (*daughter mineral*). Nel caso più generale si tratta di salgemma, silvite o anidrite. Nell'inclusione, oltre alla formazione di questi nuovi «minerali figli», si produce anche cristallizzazione lungo le pareti del cristallo ospite. Questo fenomeno, che porta a una riduzione del volume dell'inclusione originaria, è comunque di scarsissima entità e completamente invisibile.

È necessario, invece, distinguere esattamente i «minerali figli» dai «solidi» intrappolati accidentalmente. All'apparenza, questi ultimi sono del tutto simili ai minerali figli, ma la loro genesi è affatto differente. Si tratta in effetti di cristalli preesistenti alla formazione dell'inclusione rimasti poi intrappolati nell'inclusione fluida. Non derivano quindi da un processo di saturazione del fluido contenuto nell'inclusione e si riconoscono dal loro comportamento durante il riscaldamento delle inclusioni. Succede, infatti, che essi, senza subire evidenti modificazioni, persistano nelle inclusioni a temperature ben al di sopra della temperatura di omogeneizzazione, contrariamente a quanto succede per i minerali figli che a temperature prossime alla temperatura di omogeneizzazione si dissolvono.

Il problema principale a proposito della composizione delle inclusioni è la loro determinazione quantitativa. Bisogna innanzitutto considerare che, generalmente, in un campione è presente più di una generazione di inclusioni e quindi per ottenere determinazioni che abbiano un significato occorre che le misurazioni siano effettuate su singole inclusioni o gruppi di inclusioni appartenenti alla stessa generazione. Il secondo problema di rilievo concerne la possibilità concreta di contaminazioni durante il procedimento di estrazione del fluido dalle inclusioni. La maggior parte delle inclusioni ha infatti dimensioni molto piccole e contiene quindi una quantità di fluido che è ben al di sotto di quello richiesto anche con le più moderne tecniche analitiche. Milioni di piccole inclusioni contengono infatti solo pochi milligrammi di soluzione. (Molte delle analisi quantitative riportate in letteratura sono state eseguite sul campione «intero» e quindi soggette alle limitazioni sopra accennate.)

Le informazioni ottenute dalle analisi di tipo quantitativo riguardano la concentrazione salina e la composizione delle soluzioni e dei gas intrappolati entro le inclusioni. La sola proprietà che dipende dalla concentrazione salina e che è misurabile nelle inclusioni è il punto di congelamento del fluido. Quanto più alta è la salinità tanto più bassa è la temperatura di congelamento. In laboratorio questa temperatura si determina congelando dapprima completamente l'inclusione, che poi viene lentamente riscaldata mentre si tiene sotto osservazione al microscopio. La massa di cristalli solidi di ghiaccio e di sali così congelata appare al microscopio completamente opaca. Man mano che il riscaldamento procede e si raggiunge la temperatura di fusione di uno dei componenti presenti nell'inclusione, compare al microscopio una piccolissima quantità di liquido. Aumentando ulteriormente la temperatura, la fusione procede fino a quando rimarrà un ultimo cristallo di ghiaccio. La temperatura alla quale esso fonde rappresenterà la temperatura di congelamento. In teoria, quando si inverte il processo, cioè quando gradualmente si raffredda l'inclusione, dovrebbe formarsi alla stessa temperatura il primo cristallo di ghiaccio. In effetti ciò non avviene in quanto spessissimo le inclusioni devono essere raffreddate molto al di sotto del punto di congelamento prima che comincino a formarsi cristalli di ghiaccio. Questo fenomeno (sopraraffreddamento) è apparentemente imputabile al fatto che nelle inclusioni sono assenti particelle che funzionino da agenti di nucleazione di cristalli di ghiaccio. Tali particelle sono abbastanza comuni nelle acque dilavanti superficiali, in quelle meteoriche e nella neve, ma non lo sono nei fluidi delle inclusioni di origine più o meno profonda e nei fluidi che si muovono all'interno della crosta terrestre a bassissima velocità. Si ha infatti un processo di filtrazione attraverso il quale tutte le particelle presenti nei fluidi vengono più o meno eliminate prima che il fluido stesso rimanga intrappolato in un cristallo.

La situazione a proposito della determinazione della composizione delle inclusioni dovrebbe però migliorare con il progredire delle tecniche analitiche che consentiranno l'effettuazione di analisi quantitative su campioni sempre più piccoli di soluzioni. Fortunatamente si possono ottenere molte informazioni circa la composizione delle inclusioni con metodi qualitativi e semiquantitativi non distruttivi, facendo uso di un tavolino riscaldante-raffreddante applicato a un normale microscopio da petrografia.

Lo studio delle inclusioni fluide non trova applicazioni solo nella risoluzione di problemi di tipo «speculativo», ma presenta anche importantissime applicazioni «pratiche» alle quali d'altronde è connesso lo sviluppo di queste ricerche in tempi recenti.

Le inclusioni fluide sono state utilizzate soprattutto come geotermometro, ma possono costituire anche un «geobarometro» estremamente valido e dare quindi preziose informazioni sul regime di pressione esistente durante determinati processi geologici. A questo scopo sono state utilizzate, per esempio, per definire la profondità di origine dei noduli pirossenitici del Vesuvio. Le informazioni sul regime di pressione (e quindi sulla profondità) esistente al momento in cui avevano luogo determinati processi geologici sono di elevato valore per i giacimentologi, perché attraverso queste informazioni si

Nel grafico sono riportate le curve di ebollizione dell'acqua e di soluzioni saline (dallo 0 al 25 per cento di cloruro di sodio equivalente). Da queste curve è possibile ottenere la profondità assoluta di formazione delle inclusioni fluide e, quindi, dei cristalli che le contengono, a condizione che venga verificato se il fluido, al momento dell'intrappolamento, si trovava in stato di ebollizione. Altrettanto fondamentale è la determinazione della composizione delle inclusioni. Infatti dal grafico si vede come la profondità possa variare in funzione della composizione. L'acqua, per esempio, è in ebollizione in corrispondenza della isoterma di 300 gradi centigradi a 1087 metri di profondità (*A*), mentre la soluzione al 25 per cento di cloruro di sodio equivalente è in ebollizione in corrispondenza della stessa isoterma, ma a 674 metri di profondità (*B*). I dati su cui si basa il grafico sono stati ottenuti sperimentalmente da John L. Haas, Jr., dell'US Geological Survey.

Nella figura è stato sovrapposto al diagramma di fase dell'acqua pura il comportamento termico di quattro inclusioni (colonne di «capsule») aventi rapporti liquido-vapore diversi. I valori in alto indicano la pressione in atmosfere. La densità di ogni inclusione è data dal rapporto tra il volume del liquido (*in colore intenso*) a zero gradi centigradi e il volume dell'inclusione. Il rapporto liquido-vapore tende a cambiare con la temperatura e, in ogni inclusione, a una data temperatura le due fasi diventano un unico fluido omogeneo (*in grigio*). Nell'inclusione di densità 0,4 il rapporto rimane invariato, ma, alla temperatura critica (*punto sulla curva in nero continua*) la linea di separazione delle fasi (*curva tratteggiata dentro l'inclusione*) improvvisamente si affievolisce e scompare. L'inclusione è ora un fluido supercritico (ossia la sola pressione per quanto elevata non può causarne la liquefazione). La curva in nero continua separa la regione in cui esistono due fasi (*in colore chiaro*) dalla regione a una sola fase. Se nell'acqua vi sono sali in soluzione, la curva si innalza.

può stabilire l'entità dei processi erosivi che hanno interessato l'area in studio e la natura di un giacimento.

Lo studio delle inclusioni fluide trova numerose applicazioni: nella ricerca di nuovi giacimenti minerari, in particolare per meglio comprendere il loro ambiente di deposizione; nella individuazione e nella valutazione dei campi geotermici; nello studio delle fasi gassose e dei fusi silicatici presenti nelle rocce ignee; nell'identificazione di siti atti a ospitare centrali nucleari e di siti idonei per l'immagazzinamento di scorie radioattive. È chiaro che le inclusioni fluide da sole non forniscono la soluzione di tutti questi problemi, ma possono dare un aiuto decisivo quando vengano utilizzate in modo appropriato con altri dati geologici.

Passiamo ora a descrivere brevemente alcuni esempi di applicazione dello studio delle inclusioni fluide nel campo della giacimentologia, della vulcanologia, della geotermia e in problemi più strettamente applicativi, quali l'individuazione di siti per centrali nucleari e l'immagazzinamento di scorie radioattive.

Nella Mississippi Valley degli Stati Uniti vi sono giacimenti molto estesi di piombo e di zinco ospitati in calcari e dolomie di età cambriana e carbonifera e dalla località prendono il nome tutti i giacimenti di questo tipo. Le inclusioni fluide forniscono per i giacimenti piombo-zinco tipo Mississippi Valley dati che permettono di comprendere meglio l'ambiente di deposizione e che sono stati utilizzati, soprattutto da ricercatori di scuola americana, a sostegno dell'ipotesi che essi abbiano un'origine successiva alla formazione della roccia di cui fanno parte (ipotesi epigenetica), in contrapposizione alla ipotesi che si siano formati contemporaneamente alla roccia (ipotesi singenetica), sostenuta soprattutto dai giacimentologi europei. I dati ottenuti da Roedder dallo studio delle inclusioni fluide in tutti i depositi piombo-zinco tipo Mississippi Valley indicano che tali giacimenti sono stati deposti da fluidi densi, ipersalini, cioè con una concentrazione salina ben superiore alla salinità delle acque marine e che sono caratterizzati da temperature di omogeneizzazione comprese tra 75 e 200 gradi centigradi.

Secondo Roedder, se questi depositi avessero avuto origine per deposizione da acque superficiali in strutture formatesi attraverso processi carsici, le inclusioni fluide dovrebbero omogeneizzare a temperature molto più prossime a quelle dell'ambiente superficiale e dovrebbero essere molto meno saline di quanto in effetti sono. Se questi depositi avessero avuto origine per deposizione da acque marine in ambienti lagunari bassi, i fluidi dovrebbero avere caratteristiche simili alle acque marine e temperature di omogeneizzazione comprese tra 25 e 50 gradi centigradi. Anche se si supponesse che la deposizione abbia avuto luogo in ambiente costiero o continentale caratterizzato da acque salmastre e poco profonde e da forte evaporazione (sabkha) che giustificherebbe l'ipersalinità, le elevate temperature di omogeneizzazione rimarrebbero comunque incompatibili con tale tipo di ambiente. Sempre secondo Roedder i dati ottenuti dallo studio delle inclusioni fluide possono essere invece facilmente giustificati se si ammette l'esistenza di un paleoacquifero e quindi di fluidi salini connati in profondità.

Roedder ricava dalle inclusioni ulteriori informazioni sulla densità dei fluidi alle temperature di omogeneizzazione, ottenendo così indicazioni sulla paleoidrologia esistente al momento della deposizione mineraria. Poiché tale deposizione ha luogo su un lungo intervallo di tempo, anche piccole differenze nella densità dei fluidi sono più che sufficienti a far percorrere ai fluidi lunghe distanze. Questi fluidi mineralizzanti, in funzione del gradiente idraulico, seguono naturalmente il cammino con permeabilità più favorevole e di conseguenza la deposizione mineraria avviene in corrispondenza di elementi strutturali e sedimentari, quali scogliere, brecce o paleokarst (strutture originatesi per paleocarsismo). Roedder conclude che una ricristallizzazione o rimobilizzazione di precedenti depositi singenetici per mezzo di fluidi caldi e salini, come viene sostenuto da alcuni ricercatori, sembra difficile da accettare perché in questo caso si dovrebbe ammettere che non solo la ricristallizzazione (o rimobilizzazione) sia avvenuta in tutti i depositi, ma anche che tutti i depositi abbiano subito tale ricristallizzazione. Questo meccanismo dovrebbe risultare infatti più efficace a una scala molto più locale. In questo caso la scelta del modello singenetico o epigenetico ha notevoli conseguenze per quanto riguarda la fase di ricerca mineraria in termini di esplorazione, perché, in funzione del modello prescelto, possono venire individuate a priori alcune strutture geologiche quali potenziali sedi di adunamenti minerari.

Un altro tipo di depositi per il quale è particolarmente importante lo studio delle inclusioni fluide è quello che viene indicato come giacimento tipo *porphyry-copper*. Si tratta di giacimenti di grandi dimensioni, a basso tenore di rame (0,4 per cento) con mineralizzazioni disseminate a solfuri di rame e a volte con oro e molibdeno associati. Si trovano intorno ed entro corpi intrusivi ignei a composizione da dioritica a quarzo-dioritica. Gli studi sulle inclusioni fluide nei sistemi tipo *porphyry-copper* hanno dimostrato che esse derivano da uno stadio idrotermale e che sono caratterizzate da elevata salinità (>35 per cento di cloruro di sodio equivalente) e da temperature di omogeneizzazione comprese fra 250 e 700 gradi centigradi. Esistono inoltre prove che i fluidi che danno origine a questo tipo di giacimenti si trovavano in stato di ebollizione.

L'ebollizione, quindi, e l'ipersalinità, testimoniata dalla presenza di minerali figli, quali salgemma, probabilmente silvite (KCl), anidrite ($CaSO_4$) ed ematite (Fe_2O_3), sono due caratteri diagnostici di sistemi intrusivi epizonali, a cui sono normalmente associati i giacimenti tipo *porphyry-copper*. Dallo studio delle inclusioni è possibile inoltre ricavare informazioni sulla profondità a cui è avvenuta la cristallizzazione di questi corpi intrusivi ignei mineralizzati.

In fase di valutazione preliminare delle potenzialità minerarie di un determinato corpo intrusivo è molto importante poter distinguere tra un sistema tipo *porphyry-copper* e un piccolo sistema epitermale cui possono essere associate mineralizzazioni di metalli preziosi quali oro e argento e di metalli base (come piombo, zinco ecc.). I metalli preziosi e i metalli base di un sistema epitermale infatti sono depositi tipicamente da fluidi moderatamente salini, contrariamente ai depositi *porphyry-copper* che sono associati a fluidi in ebollizione, altamente salini.

In ogni fase riconoscitiva di ricerca mineraria, quindi, quando ci si trova in presenza di un corpo intrusivo «sospetto», lo studio delle inclusioni fluide va focalizzato nella ricerca di inclusioni ipersaline (con minerali figli, quale salgemma) e di prove dell'esistenza di condizioni di ebollizione. Lo studio eseguito per esempio a Bingham, Utah (USA), da Thomas Nash dell'US Geological Survey ne costituisce un esempio.

Questo tipo di ricerca si rivela fattibile anche quando ci si trova in presenza di cappellacci di alterazione di giacimenti di minerali metallici, sviluppati alla sommità di corpi intrusivi, senza la necessità di effettuare prelievi di campioni dalla roccia inalterata sottostante. È stato provato infatti che nel quarzo, anche in queste condizioni, vengono preservate le inclusioni ricche in minerali figli e che recano tracce di ebollizione, attestanti la presenza di fluidi caldi e ipersalini. Come per ogni altro tipo di prospezione, la tecnica delle inclusioni fluide diventa molto più efficace quando viene condotta in parallelo con altri studi a carattere petrologico, come per esempio lo studio delle paragenesi, della zonalità di alterazione idrotermale e degli isotopi stabili.

Solo raramente le inclusioni fluide possono essere usate per ottenere dati sia sulla temperatura sia sulla pressione, esistenti nel sistema nel momento in cui l'inclusione stessa veniva intrappolata. Il metodo più comune è quello di determinare l'isocora (linea che unisce punti con volume costante) su un diagramma pressione-temperatura e ottenere, per mezzo dell'intersezione di questa isocora con la pressione, la reale temperatura di intrappolamento, laddove la pressione si ottiene per mezzo di altri geobarometri indipendenti.

Nel caso dello studio che ho eseguito sui noduli pirossenitici del Vesuvio in collaborazione con altri ricercatori, il metodo è stato invertito utilizzando sia inclusioni magmatiche sia inclusioni contenenti anidride carbonica. Le inclusioni magmatiche si formano in minerali che cristallizzano da fusi silicatici con modalità del tutto simili alla formazione di inclusioni acquose. Come avviene nelle inclusioni acquose, in seguito a raffreddamento si possono sviluppare, per contrazione, bolle, e possono formarsi minerali figli. Poiché i fusi silicatici sono relativamente incompressibili, non si apporta alcuna

Nel grafico sono riportate la temperatura di omogeneizzazione e la salinità di inclusioni provenienti da giacimenti tipo *porphyry-copper*: nell'area in grigio sono comprese le inclusioni a due fasi (liquido-vapore) a moderata salinità; nell'area in colore le inclusioni a elevata salinità, le quali, oltre alle fasi liquido e vapore contengono «minerali figli», quali salgemma, silvite e probabilmente anidrite. La coesistenza dei due tipi di inclusione è caratteristica dei giacimenti *porphyry-copper*. I dati del grafico sono di Thomas Nash dell'US Geological Survey.

In questa cartina dell'area di Binghan Canyon negli Stati Uniti è indicata la distribuzione dei fluidi a elevata salinità (*in colore*) e i loro rapporti spaziali rispetto alla zona mineralizzata a rame (*in grigio scuro*), le zone di alterazione idrotermale e il corpo intrusivo «granitico» di Bingham (*in grigio chiaro*). Questo studio dell'area di Bingham è stato condotto da Thomas Nash dell'US Geological Survey.

L'inclusione fluida che compare in queste microfotografie dell'autore è contenuta in uno spinello derivante da un nodulo pirossenitico del periodo eruttivo del Vesuvio compreso tra il 1440 e il 1631. Nell'inclusione è presente anidride carbonica nella fase liquida e in quella gassosa. L'omogeneizzazione avviene nella fase liquida a 22 gradi centigradi. La microfotografia a sinistra mostra l'inclusione a 12 gradi centigradi, prima dell'omogeneizzazione; quella a destra è successiva all'omogeneizzazione ed è a 24 gradi centigradi. Il fatto che l'omogeneizzazione avvenga nella fase liquida ha permesso di stabilire che si tratta di un'inclusione ad alta densità di anidride carbonica (0,75 grammi per centimetro cubo). Da questo dato è possibile ricavare, come è illustrato nel grafico sottostante, la profondità di cristallizzazione dello spinello, una volta nota la temperatura alla quale questa avviene (circa 1200 gradi centigradi). Il diametro dell'inclusione è di otto micrometri.

«correzione di pressione» alla temperatura di omogeneizzazione per ottenere la temperatura di formazione. Si assume quindi che le due temperature coincidano. I metodi di studio di questo tipo di inclusioni sono simili a quelli utilizzati per lo studio delle inclusioni fluide acquose, ma poiché la temperatura di omogeneizzazione è generalmente molto più elevata (compresa tra 800 e 1200 gradi centigradi) si fa uso di una strumentazione alquanto differente. Poiché i dati sperimentali mostrano che le inclusioni vetrose sono state intrappolate insieme con la più compressibile anidride carbonica, la temperatura di intrappolamento delle inclusioni magmatiche accoppiata ai dati sulla densità dell'anidride carbonica, ottenuta attraverso la sua temperatura di omogeneizzazione, consente una stima abbastanza accurata della pressione esistente sul sistema al momento dell'intrappolamento delle inclusioni. L'anidride carbonica pura ha una temperatura critica di 31 gradi centigradi e, quindi, la fase liquida non può esistere al di sopra di questa temperatura. Se la densità dell'anidride carbonica è maggiore di 0,4 grammi per centimetro cubo, l'omogeneizzazione avverrà nella fase liquida a una temperatura inferiore a 31 gradi centigradi; se la densità è minore di 0,4 grammi per centimetro cubo, l'omogeneizzazione avverrà nella fase vapore sempre a una temperatura inferiore a 31 gradi centigradi; se invece la densità è esattamente 0,4 grammi per centimetro cubo, la bolla rimane invariata inizialmente per poi scomparire improvvisamente alla temperatura di 31 gradi centigradi (densità critica). Il fenomeno è del tutto simile a quello che abbiamo descritto nella figura di pagina 35 per l'acqua pura. Riportando quindi le informazioni così ottenute sul diagramma costruito sulla base dei dati sperimentali del ricercatore americano G. C. Kennedy, e dei sovietici V. M. Shmonov e K. I. Shmulovich, è possibile calcolare la pressione e quindi la profondità di cristallizzazione dei minerali che racchiudono le inclusioni stesse.

Nel caso specifico del Vesuvio sono stati studiati noduli pirossenitici provenienti da tre diversi periodi eruttivi (quello precedente il 79, quello tra il 1440 e il 1631 e quello del 1944). Tutti i minerali che compongono i noduli (pirosseno, olivina, spinello, apatite e biotite) contengono sia inclusioni vetrose (vetro, minerali figli e una bolla da contrazione) sia inclusioni di anidride carbonica pura (con una fase liquida, una fase gassosa più una piccola quantità di vetro).

Assumendo che la temperatura di cristallizzazione dei minerali sia di circa 1200 gradi centigradi (determinata dalla temperatura di omogeneizzazione delle inclusioni magmatiche per mezzo di un tavolino riscaldante, che raggiunge una temperatura di 1350 gradi centigradi), la densità dell'anidride carbonica (determinata otticamente dalla temperatura di omogeneizzazione dell'anidride carbonica con un tavolino riscaldante) ha permesso di determinare che parte dei noduli studiati espulsi da diversi episodi eruttivi ha subito cristallizzazione a una profondità di circa 16 chilometri, mentre un altro gruppo di noduli, più numeroso, mostra di aver cristallizzato a una profondità compresa fra tre e sette chilometri. Benché la profondità di cristallizzazione fra tre e sette chilometri fosse già nota sulla base di altri dati, il fatto di aver fissato un limite inferiore a circa 16 chilometri rappresenta un contributo nuovo, che senz'altro ha una notevole importanza per una migliore comprensione dei processi vulcanici e che, d'altra parte, dovrebbe avere ricadute di carattere applicativo circa la previsione delle eruzioni e lo studio della potenzialità geotermica del complesso Somma-Vesuvio.

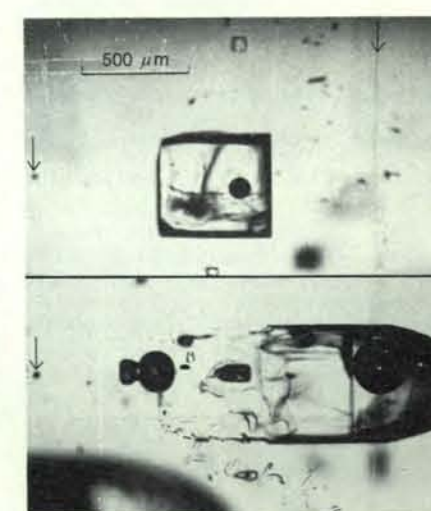
In questo diagramma, il quale mette in rapporto la temperatura e la densità per un sistema ad anidride carbonica, le isobare inferiori a 1400 bar sono state ricavate dai dati sperimentali ottenuti da G. C. Kennedy nel 1954; quelle superiori a 2000 bar si basano sui dati ottenuti dai ricercatori sovietici V. M. Shmonov e K. I. Shmulovich. I valori sull'ascissa rappresentano la densità dell'anidride carbonica determinata sperimentalmente attraverso la sua temperatura di omogeneizzazione, mentre i valori sull'ordinata indicano la temperatura di cristallizzazione dei minerali contenenti le inclusioni. La parte tratteggiata delle isobare è un'estrapolazione.

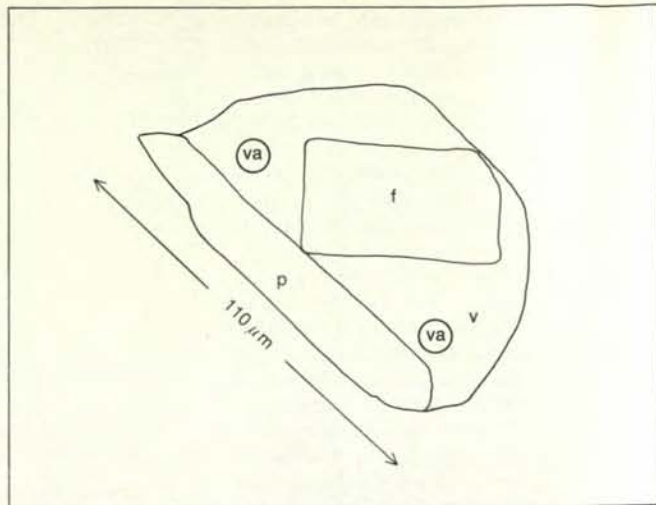
Questa inclusione magmatica in clinopirosseno, proveniente da un nodulo cumulitico di una eruzione del Vesuvio anteriore al 79 rinvenuto nella località Lagno di Pollena, contiene vetro e diverse «bolle». La presenza di molteplici bolle fa pensare che l'inclusione (fotografata dall'autore) si sia aperta durante il processo di formazione subendo una parziale perdita di fuso silicatico. Il diametro reale dell'inclusione è di circa 50 micrometri.

La tecnica delle inclusioni fluide viene applicata con sempre maggiore frequenza allo studio dell'evoluzione dei fluidi nei campi geotermici. Fornendo informazioni sulla storia passata del campo geotermico, le inclusioni fluide possono essere di utilità al fine di prevedere la vita del campo stesso e di capire l'evoluzione del regime idrologico e termico. Alcune ricerche particolarmente interessanti sono quelle condotte nel campo di Broadlands, in Nuova Zelanda, e di Larderello-Travale, in Italia.

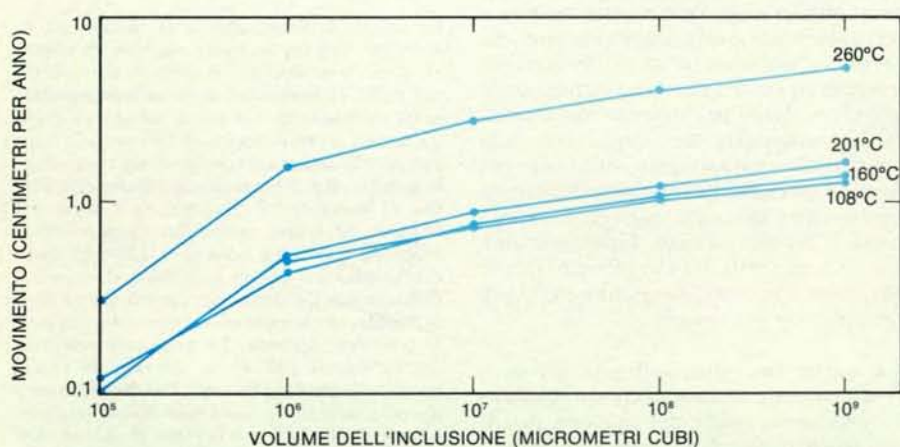
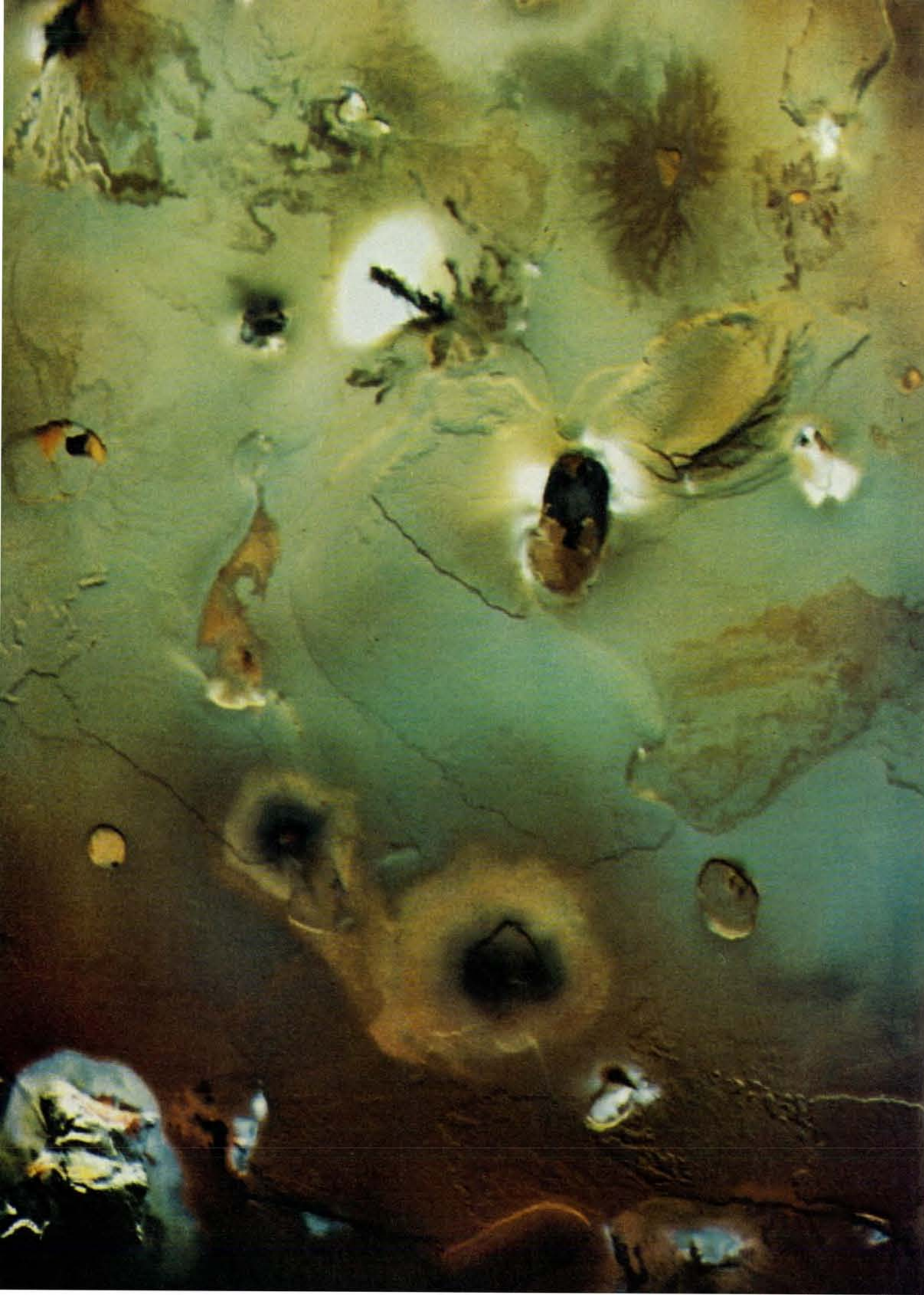
Nel campo di Broadlands, P. Browne del New Zealand Geological Survey e collaboratori (Edwin Roedder e Antoni Wodzinski) hanno trovato inclusioni entro cristalli di quarzo e blenda testimonianti che i cristalli ospiti sono stati depositati da un fluido in ebollizione. Al contrario, i liquidi presenti attualmente nei pozzi sono caratterizzati da temperature ben al di sotto del punto di ebollizione dell'acqua. La differenza fra la situazione attuale e quella registrata dalle inclusioni fluide può spiegarsi facilmente, secondo gli autori della ricerca, con una variazione del contenuto di anidride carbonica, in quanto una tale variazione può avere abbassato la curva di pressione di vapore di circa 30 gradi centigradi. Tra l'altro una diminuzione di concentrazione di anidride carbonica giustifica anche la più alta salinità apparente osservata nelle inclusioni rispetto alla salinità dei fluidi attuali. Infatti, una maggiore concentrazione di anidride carbonica determina un abbassamento del punto di congelamento e, quindi, una maggiore salinità apparente delle soluzioni contenute nelle inclusioni.

Uno studio condotto sulle inclusioni fluide del campo geotermico di Larderello-Travale da me in collaborazione con altri ricercatori ha permesso di accertare la coesistenza di inclusioni con una fase vapore e una fase liquida predominante (con salinità variabile da valori molto bassi a valori moderatamente alti) insieme a inclusioni ipersaline con minerali figli e inclusioni ricche in anidride carbonica con temperature di intrappolamento comprese tra 250 e 350 gradi centigradi. Tutto ciò testimonia l'esistenza di condizioni di ebollizione al momento dell'intrappolamento delle inclusioni. La coesistenza di fluidi con una tale variazione di salinità è perfettamente possibile, in quanto nel sistema cloruro di sodio-acqua è dimostrato sperimentalmente che a circa 300 gradi centigradi un liquido con una salinità corrispondente a una concentrazione di cloruro di sodio pari al 40 per cento può coesistere benissimo con una soluzione gassosa caratterizzata da una salinità corrispondente a una concentrazione di appena lo 0,01 per cento di cloruro di sodio. Ne deriva che tutte le inclusioni di Larderello-Travale possono essere state prodotte dall'ebollizione di acque saline: i fluidi contenenti anidride carbonica, da molto diluiti a moderatamente salini, si concentrano nella fase vapore mentre i fluidi ipersalini si originano come prodotto residuo del processo di ebollizione. La temperatura di formazione ottenuta dalle inclusioni fluide nei minerali idrotermali di Larderello-Travale, rapportata alla temperatura attualmente misurata nei pozzi, mostra che l'evoluzione del campo geotermico è avvenuta senza grosse variazioni di regime termico, fatta eccezione per un'area periferica (Val Pavone) dove potrebbe aver avuto luogo una caduta di temperatura nel tempo.

In particolari situazioni geologiche lo studio delle inclusioni fluide è un utile strumento nell'individuazione di siti «sicuri» per centrali nucleari. Nell'ambito di una indagine multidisciplinare tendente ad accertare la sicurezza, da un punto di vista geologico, del sito prescelto per la costruzione di una centrale nucleare a

La coppia di fotografie, di H. Belkin ed E. Roedder, mostra un tipico esempio di migrazione delle inclusioni ottenuto in laboratorio nel 1980. Le inclusioni sono state fotografate in un campione di sale prima (*in alto*) e dopo (*in basso*) un trattamento di 156 ore alla temperatura di 202 gradi centigradi e a un gradiente termico di 1,5 gradi centigradi per centimetro. (L'aumento di temperatura indotto dal deposito di scorie radioattive in contenitori metallici è stimato intorno a 200-250 gradi centigradi.) La grossa inclusione (*in alto*) si divide in due parti, una ricca in vapore e l'altra in liquido, che si muovono in direzione opposta al gradiente termico. La posizione originale dell'inclusione può essere rilevata da alcuni punti di riferimento (*frecce*). Per dissoluzione e riprecipitazione, le inclusioni fluide migrano attraverso il sale verso la fonte di calore, cioè verso i contenitori metallici contenenti le scorie. Con il tempo questi fluidi altamente corrosivi per il contenuto di calcio, magnesio, potassio e sodio, deteriorano i contenitori di acciaio mettendo in libertà le scorie radioattive.

Nella fotografia a sinistra, dell'autore, è visibile una inclusione magmatica in clinopirosseno proveniente da un nodulo pirossenitico della Valle dell'Inferno e che risale all'eruzione del Vesuvio del 1944. L'inclusione contiene, come è indicato nello schema a fianco, vetro (*v*), alcune bolle di vapore acqueo (*va*) (o anidride carbonica gassosa?) e grossi cristalli di «minerali figli». Il grosso cristallo rettangolare (*f*) è probabilmente flogopite, mentre il cristallo allungato (*p*) potrebbe essere un plagioclasio oppure un pirosseno. I cristalli *f* e *p* fondono rispettivamente a 1203 e a 1212 gradi centigradi. La temperatura di omogeneizzazione dell'inclusione è di 1257 gradi centigradi.

Ginna (nello Stato di New York) si è cercato di stabilire, per mezzo delle inclusioni fluide, se un sistema di faglie fosse ancora attivo e a quando risaliva il suo ultimo movimento. Per questo scopo sono stati utilizzati piccoli cristalli di calcite che si trovavano lungo il piano di faglia. Usando i dati sperimentali di John L. Haas, Jr., dell'US Geological Survey si sono ottenute informazioni sul regime di pressione esistente sul sistema nel momento in cui le inclusioni venivano intrappolate nella calcite. Poiché le variazioni di pressione occorse nella zona studiata sono funzione sia della storia glaciale che ha interessato l'area di Ginna sia dell'entità dell'erosione verificatasi nell'area, ciò ha permesso di datare con una certa approssimazione l'ultimo movimento subito dalla faglia interessata, consentendo quindi di stabilire che in termini di sicurezza il sito individuato era idoneo a ospitare una centrale nucleare. Infatti si può ragionevolmente affermare che campioni con inclusioni fluide caratterizzate da alta temperatura di omogeneizzazione e da alta salinità (e senza tracce di ebollizione) raccolti lungo faglie superficiali testimoniano un ambiente di deposizione profondo e di conseguenza che il processo erosivo è a uno stadio molto maturo. Si può dedurre quindi che la faglia è stata attiva in tempi molto remoti. Quando invece ci si trova in presenza di inclusioni con temperatura di omogeneizzazione molto bassa e con salinità molto prossima a quella delle acque sotterranee dell'area interessata, si può dire che la faglia è stata attiva in tempi molto recenti.

Come è noto, l'immagazzinamento delle scorie radioattive rappresenta forse il problema più importante connesso con la produzione di energia a mezzo di centrali nucleari. Fin dagli anni cinquanta le miniere coltivate nei duomi salini (strutture risultanti dalla risalita di masse saline) sono state considerate negli Stati Uniti siti adatti per immagazzinare scorie radioattive, dato che si suppone che queste masse saline siano ormai immobili e quasi completamente asciutte. In realtà, invece, a parte le acque provenienti da fonte esterna attraverso faglie, fratture ecc., di cui si può conoscere in anticipo, con normali indagini geologiche, l'esistenza e il comportamento, sono quasi sempre presenti anche acque interstiziali, di cui è fondamentale determinare sia la natura sia la quantità. Lo studio per mezzo delle inclusioni fluide permette di determinare quanta acqua può venire liberata e messa in movimento quando in tali depositi salini si introduce una fonte di calore esterno quale è certamente quella derivante dall'immagazzinamento di scorie radioattive. La velocità di movimento dei fluidi dipende fondamentalmente da tre fattori: temperatura ambiente, gradiente termico e dimensioni delle inclusioni. Si è dimostrato sperimentalmente che le acque interstiziali delle inclusioni si muovono verso la fonte di calore a una velocità media di circa un centimetro all'anno.

Questi dati, aggiunti a quelli sulla composizione, pressione e temperatura dei fluidi in movimento, sono fondamentali per avere un modello dell'evoluzione nel tempo di tali fluidi. La previsione del comportamento di questi fluidi interstiziali, chimicamente altamente reattivi, sotto l'influsso di una fonte di calore risulta importante in considerazione dell'effetto corrosivo che essi possono esercitare sui contenitori metallici delle scorie. I dati devono essere preliminarmente valutati con attenzione, in modo che possano essere prese le opportune contromisure tecniche in fase di progettazione ogniqualvolta viene individuato un sito per il confinamento di scorie radioattive entro duomi salini.

Il grafico mostra che nel sale il movimento delle inclusioni verso la sorgente di calore è funzione sia del volume delle inclusioni sia della temperatura. Il meccanismo del movimento è determinato dal fatto che si ha dissoluzione dal lato caldo dell'inclusione, dal lato cioè rivolto verso la fonte di calore, e precipitazione dal lato opposto. Il grafico si basa su un lavoro di H. Belkin e di E. Roedder.

Io

Eruzioni a geyser ed estese colate laviche, riprese nelle immagini delle missioni Voyager, rivelano come questo satellite di Giove sia il corpo del sistema solare che presenta l'attività vulcanica più intensa

di Torrence V. Johnson e Laurence A. Soderblom

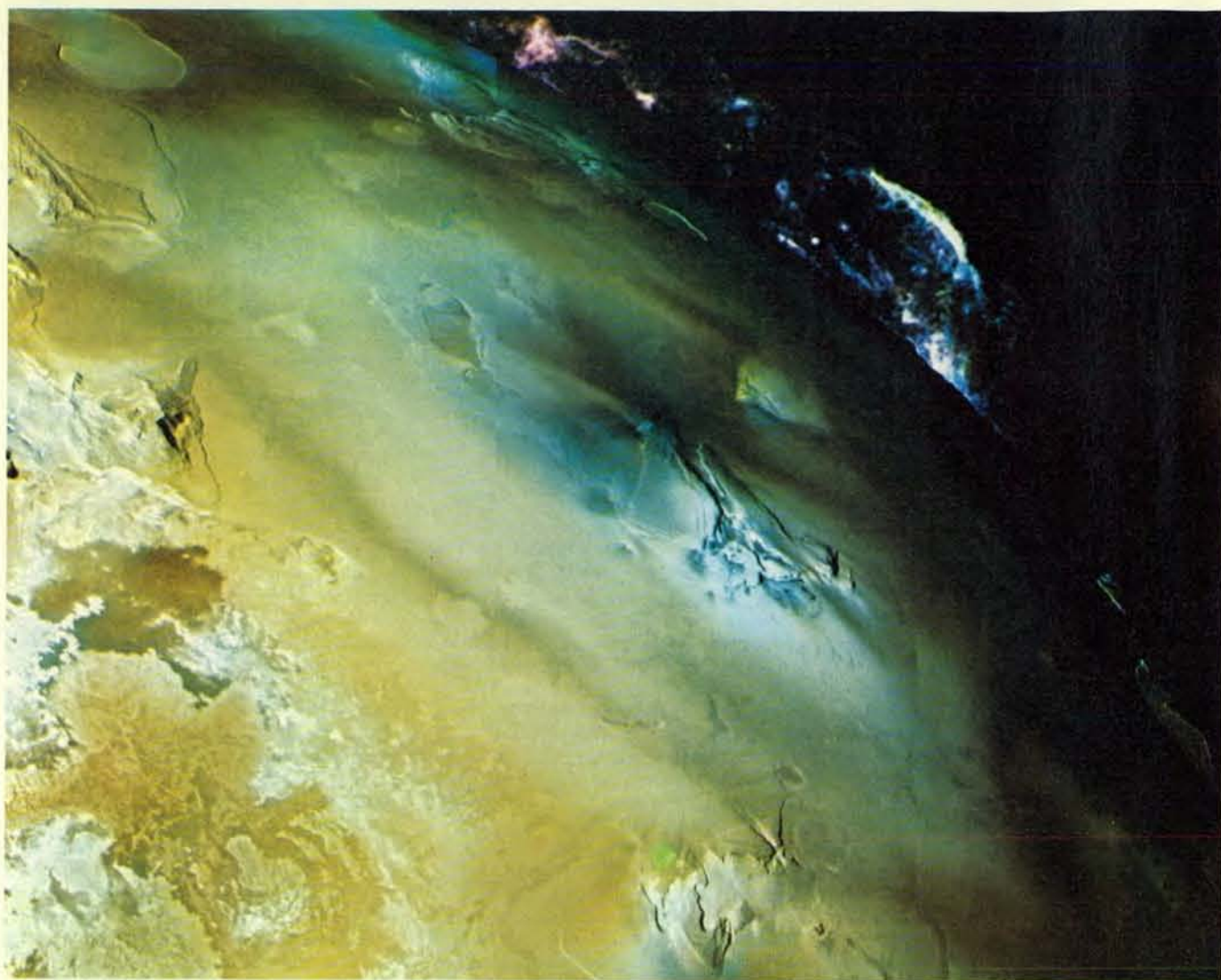
Nell'era attuale di esplorazione planetaria, Io, un satellite di Giove scoperto da Galileo quasi 400 anni fa, si è rivelato uno dei corpi più strani del sistema solare. Già negli anni sessanta gli astronomi avevano scoperto che in qualche modo Io modula gli intensi impulsi di radioonde emessi da Giove. Inoltre strumenti sensibili puntati su Io dalla Terra hanno confermato che questo oggetto ha una riflettività pari a quella della neve fresca, pur essendo di colore giallo arancione, e questa particolarità dà origine a una serie di interrogativi sconcertanti sulla composizione della sua superficie. Successivamente, negli anni settanta, si scoprì che Io inietta grandi nubi di ioni e di atomi neutri nella magnetosfera di Giove, la regione in cui le particelle elettricamente cariche vengono intrappolate dal campo magnetico del pianeta. Infine anche le sonde spaziali penetrarono nel sistema gioviano e i passaggi ravvicinati di *Voyager 1* e *Voyager 2* nel 1979 permisero di avere di Io le prime immagini ravvicinate.

Da come lo conosciamo ora, Io appare ancora più strano di prima: riscaldato dalle forze di marea esercitate da Giove, sembra il corpo più attivo dal punto di vista vulcanico di tutto il sistema solare, sicuramente più attivo della Terra. I suoi grandi geyser vulcanici scagliano pennacchi eruttivi a centinaia di chilometri di altezza e la materia che sfugge grazie a processi non ancora ben chiari dalla sua tenue atmosfera vulcanica domina veramente la magnetosfera di Giove, controllando la velocità con cui il pianeta irradia energia e influenzando probabilmente fenomeni disparati, dalle sue aurore polari agli impulsi radio provenienti dal pianeta e dai suoi satelliti. Le scoperte compiute dalle missioni Voyager hanno sollevato una serie di nuovi interrogativi; questo articolo, quindi, non può avere valore definitivo. Si tratta al contrario di un resoconto sullo studio, ancora in corso, di un oggetto celeste notevole.

Io è uno dei satelliti di Giove «normali», quelli che percorrono orbite sostanzialmente circolari all'incirca sullo stesso piano dell'equatore del pianeta e che, quindi, si crede si siano formati intorno a Giove in un processo molto simile a quello di formazione dei pianeti attorno al Sole. A una distanza orbitale di 350 000 chilometri (circa uguale a quella tra Terra e Luna), Io è il più interno dei quattro satelliti maggiori di Giove, quelli scoperti da Galileo. Io ed Europa, il successivo del gruppo in ordine di distanza, hanno all'incirca le dimensioni del nostro satellite e una densità simile a quella delle rocce terrestri: rispettivamente 3,5 e 3 grammi per centimetro cubo. Ganimede e Callisto, i due satelliti più esterni, sono più grandi (circa come Mercurio), hanno una densità decisamente più bassa, intorno ai due grammi per centimetro cubo, e sembrano costituiti per metà di ghiaccio d'acqua e per metà di roccia, soprattutto silicati. La progressiva variazione di densità dall'uno all'altro dei satelliti è stata spesso considerata una tendenza regolare, ma sembra che in realtà segua una distribuzione bimodale. La massa dedotta per i silicati di Ganimede e di Callisto è circa pari alla massa totale di Io o di Europa; si può pensare, quindi, che i quattro corpi siano simili, ma che i due più interni stiano perdendo la loro quota di sostanze volatili, soprattutto acqua. (Si sa che Europa è coperto di ghiaccio, ma la sua densità complessiva indica che questa copertura non è che uno strato molto sottile, di spessore inferiore a un ventesimo del raggio del satellite.)

La variazione di densità è attribuita di solito all'effetto che Giove ha avuto sulle condizioni primitive nelle sue vicinanze. Ancora oggi alle lunghezze d'onda infrarosse Giove irradia quasi il doppio dell'energia che riceve dal Sole; è chiaro che sta raffreddandosi lentamente da uno stato originario luminoso. Per questo motivo i modelli delle prime fasi della sua storia (basati su una fisica molto simile a quella delle stelle di piccola massa) fanno credere che il pianeta primitivo sia stato un'importante sorgente locale di energia per diverse centinaia di milioni di anni. I modelli sviluppati da James B. Pollack dell'Ames Research Center della National Aeronautics and Space Administration e dai suoi collaboratori fanno pensare che per un periodo dell'ordine dei cento milioni di anni la temperatura e la pressione nella posizione attualmente occupata dall'orbita di Io siano state determinate soprattutto da Giove e non dal Sole. In realtà, nel primo milione di anni di quella fase, un corpo nell'orbita di Io avrebbe ricevuto da Giove più energia di quanta la Terra non ne riceve oggi dal Sole.

È probabile, quindi, che la composizione dei satelliti normali di Giove sia stata influenzata dalla loro maggiore o minore distanza dal pianeta. In questa prospettiva Io ed Europa sono corpi rocciosi perché le temperature nelle vicinanze di Giove erano troppo elevate per permettere a questi satelliti di trattenere quantità significative di acqua durante la loro formazione. Ganimede e Callisto, invece,

L'emisfero meridionale di Io è stato fotografato nel 1979 con un grandangolare a bordo di *Voyager 1*. Il campo di vista di questa immagine abbraccia circa 2,5 milioni di chilometri quadrati, cioè il 6 per cento della superficie di Io; un centimetro rappresenta circa 50 chilometri. Si vedono diverse classi di strutture: le più numerose sono le caldere (crateri dovuti al collasso di vulcani). Alcune sono circondate da caratteristiche di grande estensione che probabilmente rappresentano la colorazione lasciata sulla superficie dai geyser solforosi. Altre sono circondate da formazioni che probabilmente sono colate eruttive, e altre ancora sembrano piene di materiale eruttivo, per esempio la caldera di forma romboidale con le macchie rosso-nere. Si pensa che le chiazze bianche che costellano la superficie siano depositi di miscele di gas e brina d'anidride solforosa eruttate lungo alcune scarpate e fratture. La montagna che si può osservare in basso a sinistra è ritenuta un affioramento della crosta silicatica di Io, e si calcola che sia alta circa 10 000 metri. L'immagine è stata elaborata da un calcolatore dell'US Geological Survey in modo che le diverse strutture fossero visibili come se l'osservatore si trovasse direttamente sopra il panorama.

Pele è stato il primo vulcano scoperto su Io ed è anche la più grande eruzione analoga a un geyser finora osservata sul satellite. Il pennacchio del geyser, visibile sopra il bordo del satellite, raggiunge un'altezza di 300 chilometri e ha depositato sulla superficie una serie di anelli concentrici di colore giallo e marrone: il più esterno ha un diametro medio di 1400 chilometri. La sorgente del geyser, al centro del deposito, è un gruppo di colline con una valle centrale. Sono visibili anche tracce e colate di eruzioni precedenti. Questa vista di Pele, ripresa da *Voyager 1*, è un mosaico ottenuto con una tecnica elaborata da Alfred S. McEwen del Geological Survey in cui le immagini ad alta risoluzione danno i dettagli spaziali e quelle a bassa risoluzione i colori. All'arrivo di *Voyager 2*, quattro mesi dopo *Voyager 1*, Pele era inattivo.

che conservarono una certa quantità d'acqua, divennero miscugli di roccia e ghiaccio e raggiunsero dimensioni maggiori. È notevole che le loro densità siano all'incirca quelle previste nel caso che un gas con la composizione del Sole venisse raffreddato fino alla temperatura di condensazione del ghiaccio.

Un quadro così semplice non può esaurire l'argomento: in primo luogo, tutti i corpi del sistema solare sono stati soggetti a una pioggia di craterizzazione molto intensa da parte di planetesimi in un periodo di pesante bombardamento, terminato circa quattro miliardi di anni fa. I planetesimi che cadevano nel campo gravitazionale di Giove subivano una forte accelerazione e quindi può darsi che Io ed Europa abbiano risentito di questo effetto molto più dei satelliti più lontani. Inoltre l'intensa attività vulcanica di Io ne ha certamente modificato il contenuto di sostanze volatili: in più, sembra che Io, e più in generale i satelliti normali di Giove, siano derivati da processi analoghi a quelli che hanno dato vita ai pianeti. Giove e i suoi satelliti sono quindi un sistema solare in miniatura in un senso più profondo di quello meramente geometrico.

La superficie di Io così come appare nelle immagini riprese dai Voyager è dominata dall'attività vulcanica in misura quasi del tutto inattesa, ed eccone la ragione. Si ritiene che i fatti che influenzano normalmente l'evoluzione termica di un corpo planetario, e quindi il suo vulcanismo, siano, in primo luogo, il riscaldamento dovuto alla liberazione di energia potenziale gravitazionale nel corso del suo accrescimento e anche al decadimento degli isotopi radioattivi a vita breve come l'alluminio 26; in secondo luogo, l'accumulo graduale di calore dovuto al decadimento degli isotopi radioattivi a vita lunga, principalmente quelli dell'uranio, del torio e del potassio; infine il raffreddamento progressivo che ha luogo via via che le fonti di calore radioattive si esauriscono e il calore si disperde attraverso la superficie del corpo per convezione e per conduzione. È una successione di fenomeni che favorisce, come sedi di attività vulcanica, i grandi corpi solidi in cui il rapporto tra volume e area superficiale è più elevato che in quelli più piccoli. Per questo motivo la Terra oggi è piuttosto attiva, mentre i centri di attività vulcanici sulla Luna sono estinti da lungo tempo. (Quelli di Marte sono probabilmente ancora attivi, ma molto più deboli di quelli della Terra.) Dato che Io ha le dimensioni e la massa della Luna, non dovrebbe mostrare altro che tracce di un antico vulcanismo scomparso.

Questa previsione cambiò improvvisamente e in maniera sensazionale nel marzo 1979 solo qualche giorno prima che *Voyager 1* raggiungesse il sistema gioviano, quando Stanton J. Peale dell'Università della California a Santa Barbara e Ray T. Reynolds e Patrick M. Cassen dell'Ames Research Center pubblicarono un'analisi da cui risultava che le forze di marea erano alla base del processo di riscaldamento di Io. Io ed Europa sono imprigionati in una delle configurazioni di risonanza gravitazionale studiate per la prima volta da Pierre Simon de Laplace all'inizio del XIX secolo. In una situazione del genere l'interazione gravitazionale di due satelliti modifica le loro distanze orbitali finché i loro periodi orbitali non sono multipli l'uno dell'altro. (Il periodo orbitale di Europa è il doppio di quello di Io.) Un effetto della risonanza Io-Europa è che Europa introduce continuamente una certa eccentricità nell'orbita di Io, cioè la devia dalla forma circolare. Di conseguenza il punto di Io più vicino a Giove oscilla, producendo un rigonfiamento mareale che si sposta avanti e indietro sulla superficie di Io e varia di ampiezza riscaldando per attrito il satellite. Peale e collaboratori calcolarono che la quantità di calore prodotta avrebbe potuto essere sorprendente e ipotizzarono la presenza di caratteristiche vulcaniche molto marcate.

Durante l'avvicinamento di *Voyager 1* a Io apparve evidente che Peale e i suoi collaboratori avevano ragione. Nelle immagini di risoluzione via via crescente non comparivano tracce di crateri da impatto di grande scala. La superficie era invece ricoperta di chiazze multicolori gialle, arancione e rosse, che si risolsero ben presto in strutture simili a colate laviche e a caldere (crateri dovuti al collasso di vulcani) terrestri e marziane. Fu ovvio che Io ha una superficie attiva, geologicamente giovane. Qualche giorno dopo il passaggio ravvicinato di *Voyager 1* altre analisi delle immagini condotte dal gruppo addetto alla navigazione e da quello addetto alla tecnica di produzione delle immagini della missione Voyager rivelarono l'esistenza di eruzioni vulcaniche in corso. Da allora la natura dell'attività vulcanica rilevata dai Voyager è stata oggetto di intensi studi, dai quali comincia a emergere un quadro preliminare della cinetica e della termodinamica del vulcanismo di Io.

Molte caratteristiche dei vulcani di Io derivano dalla chimica eccezionale della sua superficie. Già molto prima del passaggio di *Voyager 1* l'assenza di certe caratteristiche nello spettro infrarosso di Io (in particolare righe di assorbimento dovute a ghiaccio o a brina d'acqua) aveva rivelato che la superficie di Io è estremamente secca e alcuni dettagli del suo spettro alle lunghezze d'onda del visibile avevano fatto supporre che vi si trovasse zolfo sotto qualche forma: anche le righe di emissione spettrale dovute agli ioni zolfo imprigionati nella magnetosfera di Giove costituivano un indizio in questo senso. In seguito tutta una serie di strumenti a bordo di *Voyager 1* e *Voyager 2* rivelò che gli ioni zolfo e ossigeno sono diffusi in tutta la magnetosfera, ma sono concentrati nelle vicinanze dell'orbita di Io, e che le radiazioni ultraviolette di questi ioni provengono da un toro intorno a Giove centrato sull'orbita di Io. Infine i membri del gruppo di lavoro addetto allo spettrometro infrarosso delle missioni Voyager scoprirono alcune caratteristiche dello spettro di assorbimento infrarosso dovute ad anidride solforosa gassosa al di sopra di uno dei vulcani del satellite. Questo chiarì che lo zolfo e i suoi composti rivestono un ruolo fondamentale nell'attività vulcanica di Io. Spronati dalle scoperte dei Voyager due gruppi di ricerca, uno al Jet Propulsion Laboratory del California Institute of Technology e uno all'Università di Hawaii, scoprirono che una caratteristica evidente nello spettro infrarosso di Io a una lunghezza d'onda di circa 4,1 micrometri era dovuta a brina di anidride solforosa.

Perché lo zolfo è così importante su Io? Nell'universo questo elemento è abbastanza comune; è soltanto 40 volte meno abbondante dell'ossigeno. Nella crosta terrestre invece è scarso, ma questo fa credere che lo zolfo terrestre sia in gran parte nascosto nel nucleo sotto forma di solfuro di ferro. Anche sulla Luna lo zolfo è scarso, ma in questo caso la sua penuria è parte di un problema più vasto: il basso contenuto di composti volatili del nostro satellite, che rimane uno dei maggiori ostacoli per qualsiasi tentativo di spiegarne l'origine. Nelle meteoriti, soprattutto in quelle primitive, lo zolfo è abbondante, e per quanto riguarda la superficie di Marte i dati raccolti dai moduli di atterraggio dei Viking rivelano una abbondante presenza di zolfo in varie forme.

Quasi tutti i modelli relativi alla formazione dei satelliti di Giove indicano che inizialmente Io doveva essere più ricco di sostanze volatili della Luna. Bisogna supporre, quindi, che il forte riscaldamento mareale nel corso della storia del sistema solare abbia allontanato tutte le sostanze volatili leggere presenti su Io, compresa l'acqua, e che quindi quelli dello zolfo siano gli unici composti volatili rimasti in grandi quantità. Un'alternativa possibile è che la temperatura nelle vicinanze di Io all'inizio della sua formazione sia stata abbastanza elevata da impedire l'inclusione di acqua, ma non abbastanza da evitare quella di zolfo. La natura di Amaltea, il satellite immediatamente più vicino a Giove, potrebbe risolvere la questione: se Io fosse stato troppo caldo per trattenere l'acqua già all'inizio della sua formazione, Amaltea avrebbe dovuto esserlo ancora di più (supponendo che si sia formato nella posizione che occupa oggi rispetto agli altri satelliti gioviani) e quindi dovrebbe essere costituito di materia estremamente refrattaria. Sulla base di dati spettrali abbastanza scarsi, però, è stata avanzata l'ipotesi che la superficie di Amaltea sia analoga alle meteoriti primitive, le condriti carbonacee, e che pertanto contenga quantità significative di sostanze volatili.

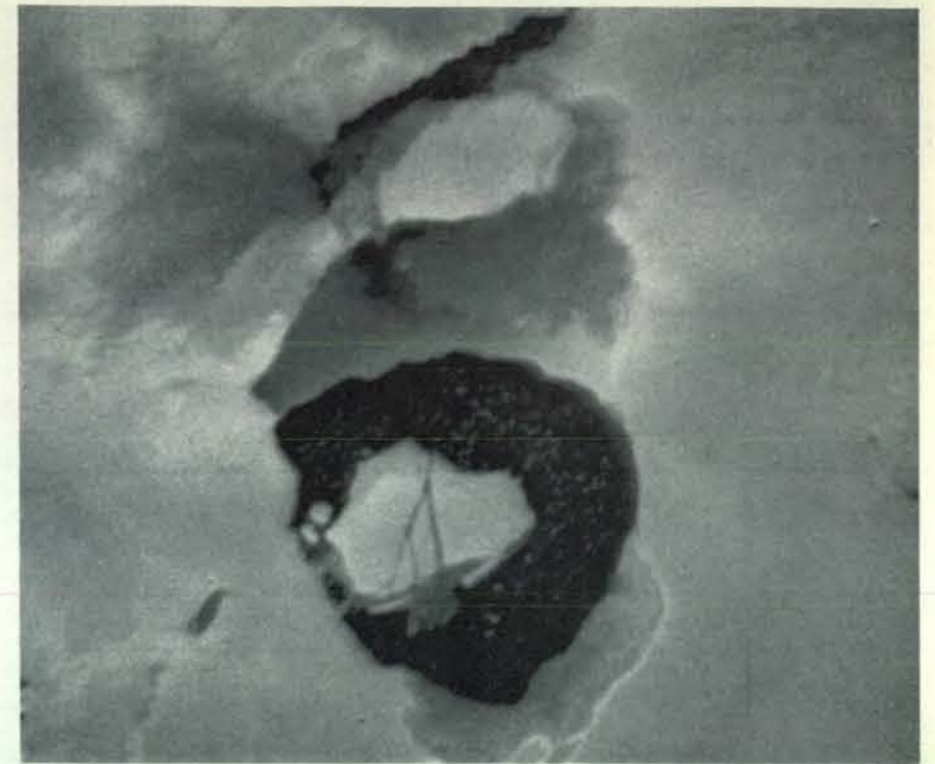
Tra i tipi di attività vulcanica presenti su Io va fatta una distinzione; tutti probabilmente sono influenzati dalla chimica dello zolfo. A un tipo, quello dei pennacchi eruttivi, va fatta risalire la forma di attività vulcanica più spettacolare osservata fino a oggi sul satellite. Si consideri Pele, attivo durante il passaggio ravvicinato di *Voyager 1*, il quale è stato il primo pennacchio a essere scoperto. Pele si innalzava fino a 300 chilometri di altezza, distribuendo il materiale eruttato in una struttura a ombrello di circa 1400 chilo-

Prometeo rappresenta una classe di pennacchi vulcanici più piccoli, più freddi e di maggior durata di quelli che assomigliano a Pele. Queste tre immagini di Prometeo sono state ottenute da *Voyager 1*: quella in alto è una vista di profilo del pennacchio, alto 100 chilometri e largo 300, le altre mostrano il pennacchio mentre la sonda lo sorvola. Il pennacchio appare scuro contro lo sfondo della superficie, ma ha depositato un anello di materiale chiaro. Il chiarore indica che si tratta probabilmente di accumulo di anidride solforosa congelata.

I pennacchi Loki spuntano dalle estremità di una fessura diritta e nera lunga circa 200 chilometri. Entrambi hanno emesso un «ventaglio» chiaro e un deposito scuro più grande, come se combinassero le due forme di eruzione di Prometeo e di Pele. Sotto la fessura Loki c'è una chiazza nera a forma di *D* che probabilmente è un lago di lava; è ancora da chiarire però se si tratti di zolfo o di silicati. Le misurazioni dell'emissione infrarossa di Io eseguite dai Voyager indicano che quasi tutta la superficie del lago ha una temperatura di circa 300 kelvin, circa 170 gradi più elevata di quella del terreno circostante. L'immagine è un mosaico fotografico di *Voyager 1* elaborato con la tecnica di McEwen.

metri di diametro. La forma a ombrello che contraddistingue molti pennacchi indica che il materiale espulso viaggia su traiettorie balistiche. D'altra parte, però, una serie di fotografie del pennacchio di Pele rivela una struttura vorticosa. I vortici fanno pensare a una turbolenza e, quindi, a un'interazione tra particelle visibili e un qualche gas invisibile che le trascina. (Nelle immagini riprese dai Voyager sono visibili soprattutto particelle solide di dimensioni micrometriche o inferiori; i gas a esse associati non sarebbero visibili direttamente.) I modelli al calcolatore di pennacchi formati da particelle eruttate balisticamente nelle condizioni gravitazionali di Io si accordano con le caratteristiche generali dei pennacchi simmetrici di Io.

Qualunque modello dei meccanismi responsabili dei pennacchi deve spiegare le alte velocità di uscita ricavate dall'analisi balistica, che vanno dai 500 ai 1000 metri al secondo. Bradford A. Smith dell'Università dell'Arizona e i suoi collaboratori Eugene M. Shoemaker e Susan W. Kieffer dello US Geological Survey e Allan F. Cook II del Center for Astrophysics dello Harvard College Observatory e dello Smithsonian Astrophysical Observatory hanno proposto un modello a geyser, basato sull'ipotesi che la grande quantità di calore che le maree «pompano» all'interno di Io debba tradursi, al di sotto della superficie, in temperature tali da far entrare in contatto a profondità relativamente ridotte l'anidride solforosa liquida con lo zolfo fuso. Usando questa ipotesi Susan W. Kieffer ha compiuto uno studio approfondito della termodinamica delle eruzioni a pennacchio che coinvolgono queste sostanze su una vasta gamma di condizioni. L'estremità a bassa energia di una gamma di eruzioni possibili è l'eruzione che inizia quando l'anidride solforosa liquida entra in contatto con zolfo liquido caldo. L'anidride solforosa liquida inizia a bollire cosicché una miscela di liquido e gas comincia a espandersi verso la superficie attraverso un qualche condotto. Mentre l'espansione procede, la produzione di vapore aumenta. Poi l'anidride solforosa raggiunge il suo punto triplo, ossia la temperatura alla quale coesistono le fasi solida, liquida e gassosa. A questo punto, ancora a una certa profondità al di sotto della superficie di Io, il liquido residuo congela e l'ulteriore espansione dell'anidride solforosa è evidenziata dalla condensazione di vapore di anidride solforosa in una «neve» di particelle solide. Un'alternativa a energia superiore ha inizio da vapore di anidride solforosa surriscaldata. In questo caso durante l'espansione verso la superficie si verificano prima la condensazione del liquido e poi il congelamento e la neve.

In ogni caso la miscela delle diverse fasi di anidride solforosa viene accelerata entro il condotto che la porta in superficie fino alla velocità del suono in quel mezzo (qualche centinaio di metri al secondo) ma non può superarla. Più vicino alla superficie può darsi che il condotto si allarghi. Inoltre, quando

Questa immagine, sempre ottenuta da *Voyager 1*, è un primo piano del lago Loki. La grande «zattera» chiara sul lago è percorsa da diverse crepe, e numerosi pezzi di materiale chiaro sparsi per tutta la superficie del lago sembrano essersi staccati dai suoi bordi. Forse la crosta del lago in via di raffreddamento è stata frantumata dai moti convettivi o da un'aggiunta di nuovo materiale eruttivo. Si può pensare che la zattera sia di zolfo elementare. Il lago è largo 250 chilometri.

raggiunge la superficie, il flusso può espandersi nell'atmosfera rarefatta di Io e quindi può raggiungere sopra la bocca del condotto velocità impressionanti. Calcolando la quantità di energia disponibile, la Kieffer ha stimato che sono possibili velocità di uscita superiori a 1000 metri al secondo, facendo però notare che le osservazioni fatte dai Voyager possono essere spiegate da una vasta gamma di condizioni al di sotto della superficie e che quindi non è possibile specificare per i singoli pennacchi i valori della temperatura, della profondità di origine e della composizione.

Si possono ricavare ulteriori informazioni sui pennacchi dalla loro distribuzione sulla superficie di Io e dal loro comportamento nel tempo. All'arrivo di *Voyager 1* erano visibili almeno nove pennacchi in eruzione, di cui alcuni vennero osservati diverse volte durante i passaggi. I nove pennacchi erano distribuiti in maniera più o meno uniforme in longitudine, ma concentrati alle basse latitudini: otto di essi si trovavano entro 30 gradi dall'equatore. Quattro mesi dopo, all'arrivo di *Voyager 2*, di questi nove pennacchi solo Pele, il più grande, non era più attivo. Due pennacchi formavano una coppia, ognuno a un'estremità di una formazione a fessura a nord di una macchia nera chiamata Loki. Durante i passaggi ravvicinati e anche tra un passaggio e l'altro, i due pennacchi rivelarono un comportamento eruttivo molto variabile in altezza. Inoltre le caratteristiche della superficie intorno alle bocche Loki e Pele sono cambiate radicalmente durante i due passaggi.

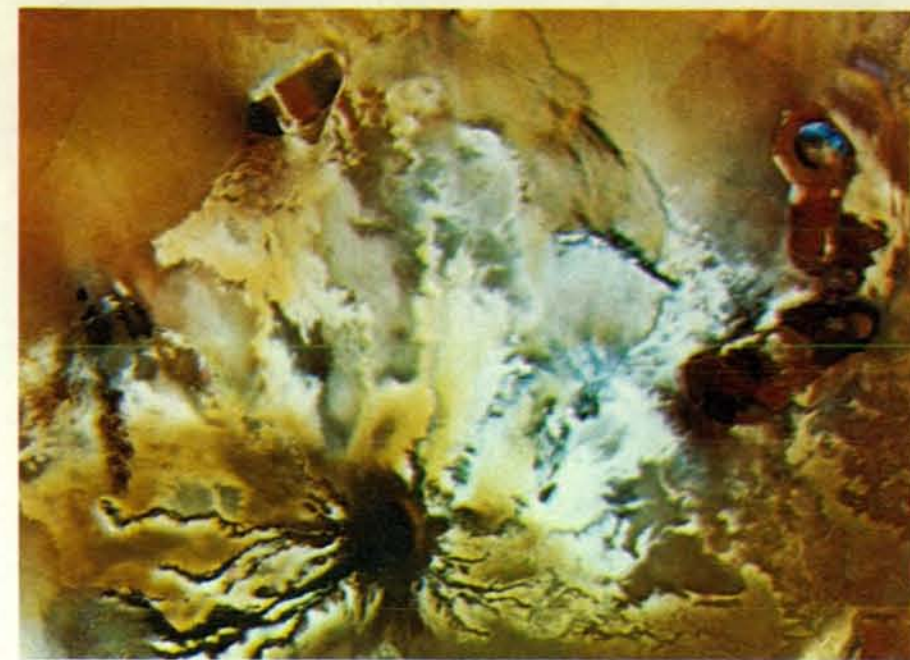
Pele e Loki si trovano abbastanza vicini l'uno all'altro. In particolare si trovano entrambi in un emisfero intermedio tra quello rivolto verso Giove e quello «nascosto». (Io, come la Luna, volge verso il proprio pianeta sempre la stessa faccia.) Da tempo si sa che questo emisfero è in media molto più rosso e più scuro dell'altro. Anche i soli altri mutamenti di rilievo avvenuti sulla superficie di Io tra i due passaggi hanno avuto luogo qui. Attorno a due caldere situate a latitudini elevate, una a nord, Surt, e una a sud, Aten Patera, si sono sviluppati due depositi scuri rossastri, a forma di anello, di circa 1400 chilometri di diametro, cioè circa le dimensioni del deposito di Pele. Nessuna delle due fu vista in eruzione durante il primo passaggio dei Voyager, ma forse l'attività di Surt è stata la causa del periodo di maggior emissione termica alla lunghezza d'onda di cinque micrometri (nell'infrarosso) osservata una notte al Mauna Kea Observatory da William M. Sinton dell'Università di Hawaii a Honolulu circa un mese prima del passaggio di *Voyager 2*. Il deposito di Aten Patera è stato scoperto in una analisi delle immagini dei Voyager effettuata ultimamente da Alfred S. McEwen e da uno di noi (Soderblom) dello US Geological Survey.

Maasaw Patera e le caratteristiche che circondano questa caldera di Io sono molto simili a quelle che i geologi incontrano sulla Terra e su Marte. La differenza principale è che Maasaw Patera è enorme: la sola caldera è larga 50 chilometri. I due livelli di Maasaw Patera fanno pensare a diversi stadi di un collasso: la parte più estesa della caldera, secondo i calcoli basati sulle ombre al suo interno, è profonda 700 metri, mentre il resto, la regione in alto a destra, è profondo 2000 metri. In alcuni punti la lava, sgorgando, deve aver superato i bordi, formando colate lunghe centinaia di chilometri. Questa immagine è un altro dei mosaici fotografici di *Voyager 1* elaborati da McEwen.

La colorazione e le dimensioni di questi enormi depositi li rendono molto simili a quello che circonda Pele; per questo McEwen e Soderblom avanzano l'ipotesi che essi appartengano a una classe di grandi eruzioni intense di breve durata di cui Pele è il miglior esempio finora noto. Utilizzando misurazioni fatte dallo spettrometro infrarosso di *Voyager 1*, John C. Pearl e collaboratori del Goddard Space Flight Center della National Aeronautics and Space Administration valutano che la temperatura della bocca Pele sia di circa 650 kelvin; le osservazioni di Sinton, che presumibilmente riguardano Surt, suggeriscono un valore analogo.

Esiste una seconda classe di eruzioni, più piccole, che depositano anelli di circa 300 chilometri di diametro. Si tratta di fenomeni di lunga durata, osservati tutti da entrambi i Voyager. Da questo fatto si può stimare che la loro durata sia di qualche anno almeno. Sembra anche che siano più freddi: circa 400 kelvin. Queste eruzioni sono limitate a una fascia equatoriale di materiale bianco luminoso, che si ritiene ricco di anidride solforosa congelata. (I loro depositi ad anello sono formati dallo stesso materiale.) L'esempio migliore è la struttura chiamata Prometeo, cinque gradi a sud dell'equatore. La coppia di pennacchi alle estremità della fessura Loki sembrava oscillare tra le due classi. I depositi che circondano questi due geyser sono complessi: ciascuno ha una zona interna costituita da un deposito a forma di ventaglio, probabilmente ricco di anidride solforosa, lungo circa 200 chilometri, e una zona esterna simile per dimensioni e colore ai depositi che si trovano intorno ai tre grandi pennacchi di tipo Pele.

McEwen e Soderblom suggeriscono che delle due classi di eruzioni siano responsabili due sistemi vulcanici ben diversi, alimentati da sostanze volatili differenti, e che la diversità delle classi derivi dalla particolare relazione esistente nello zolfo tra temperatura e viscosità. Quando viene riscaldato al di sopra del suo punto di fusione, infatti, lo zolfo subisce numerose trasformazioni successive. All'inizio è un solido giallo; poi, a una temperatura di circa 400 kelvin, fonde e forma un liquido giallo a bassa viscosità. Un ulteriore riscaldamento trasforma il suo colore prima in arancione e poi, a 430 kelvin, in rosa. Successivamente diventa rosso vivo e più denso; verso i 500 gradi circa si trasforma in un catrame nerastro. A circa 600 kelvin la sua viscosità comincia a diminuire e prima dei 650 kelvin torna interamente fluido. Infine, a una temperatura più elevata che dipende dalla pressione, il liquido si trasforma in vapore.

Lo zolfo fuso, poco viscoso, in grado di trasportare facilmente il calore esiste, quindi, in due forme: una trasparente rossastra tra i 400 e i 430 kelvin e una nera opaca sopra i 650 kelvin. L'idea, quindi, è che i pennacchi piccoli come Prometeo siano alimentati da zolfo liquido rosso che entra in contatto con anidride solforosa liquida e la fa espandere ed evaporare

Ra Patera e le caratteristiche superficiali circostanti sono molto diverse da quello che i geologi osservano sulla Terra e su Marte. Qui Ra Patera è nella parte inferiore dell'immagine. Si tratta di una caldera piena di materiale nero che deborda verso l'esterno formando lunghe colate sinuose, soprattutto verso sinistra, che si estendono fino a 200 chilometri di distanza. David C. Pieri del Jet Propulsion Laboratory del California Institute of Technology e Carl Sagan della Cornell University hanno avanzato l'ipotesi che il centro scuro di Ra Patera sia costituito da una fase dello zolfo nera e catramosa, mentre le colate sarebbero costituite da fasi più fredde e meno viscose. Nell'immagine si vedono altre due caldere: quella allungata in alto a sinistra è a strisce gialle, arancione e nere; i colori forse corrispondono a diverse fasi dello zolfo. In alto a destra, invece, c'è una caldera circolare che contiene una mezzaluna azzurra, che in una fotografia ripresa da *Voyager 1* sei ore prima di questa non c'era. Forse si tratta del segno di un'eruzione a bassa energia.

come sostengono Smith, Shoemaker, Kieffer e Cook, mentre i pennacchi grandi come Pele hanno origine quando i silicati caldi della crosta di Io vaporizzano lo zolfo a temperature comprese tra 700 e 1200 kelvin.

La notevole differenza di durata tra le due classi di pennacchi si può spiegare con questo modello: i pennacchi piccoli di anidride solforosa hanno una durata di vari anni perché l'anidride solforosa può rimanere liquida in un esteso intervallo di profondità all'interno della crosta di Io; il liquido ha una viscosità molto bassa (all'incirca come l'alcool a temperatura ambiente) e quindi migra facilmente nella crosta per raggiungere le bocche in superficie. I grandi pennacchi di zolfo nero, invece, durano solo qualche giorno perché le condizioni che permettono la loro eruzione hanno un equilibrio molto delicato. Il flusso dello zolfo potrebbe fermarsi, per esempio, se alla bocca di uscita, nel condotto, o anche nelle regioni più profonde della crosta, dove l'elemento si incontra con i silicati, la temperatura scendesse molto al di sotto dei 650 kelvin perché lo zolfo condenserebbe rapidamente in una sostanza semisolida simile alla pece.

La spiegazione della distribuzione spaziale delle due classi di pennacchi è ancora un problema aperto. Le immagini ottenute dai Voyager dell'emisfero rosso scuro di Io, dove si trovano pennacchi grandi, mostrano numerose montagne, composte con tutta probabilità di silicati, che si elevano al di sopra delle pianure di Io. (Una montagna di zolfo infatti non sarebbe abbastanza resistente da sostenere il proprio peso.) Le immagini dell'emisfero opposto hanno una risoluzione inferiore, ma in alcune di esse la regione in cui sono più abbondanti i piccoli pennacchi simili a Prometeo si trova sul terminatore (la linea di confine tra notte e giorno), lungo la quale le montagne dovrebbero essere facilmente visibili, mentre in altre si trova sul bordo illuminato di Io. Tutte queste inquadrature mostrano soltanto pianure uniformi.

Sulla base di queste osservazioni McEwen e Soderblom ipotizzano che la crosta del satellite vari in misura notevole da un punto all'altro e che in particolare sia molto più sottile che altrove nella regione dei grandi pennacchi simili a Pele, cosicché i silicati ad alta temperatura si troverebbero più vicini alla superficie e la pressione che confina il vapore di zolfo sarebbe inferiore. Inoltre la crosta della regione dei piccoli pennacchi simili a Prometeo potrebbe essere ricca di anidride solforosa liquida a causa di qualche processo geofisico (per esempio un gradiente di pressione idrostatica variabile con la latitudine e generato dalla rotazione di Io) che spingerebbe il liquido verso l'equatore. Vale la pena di osservare che la coppia di pennacchi chiamata Loki, forse un

ibrido tra le due classi di pennacchi, si trova in una zona in cui le regioni delle due classi si sovrappongono.

Una forma di attività vulcanica molto diversa da quella dei pennacchi analoghi ai geyser è quella delle grandi caldere vulcaniche, delle colate e delle caratteristiche della superficie a esse associate. Si tratta delle strutture più evidenti identificate nelle immagini dei Voyager. Di solito le caldere assomigliano a quelle terrestri, ma hanno dimensioni maggiori: più di 200 sono larghe almeno 20 chilometri, contro le 15 della Terra la cui superficie continentale è maggiore di 3,5 volte. Contrariamente ai pennacchi, le caldere di Io non mostrano una particolare concentrazione in latitudine, avendo quasi la stessa abbondanza per unità di area sia alle alte latitudini sia all'equatore.

Il dibattito sulla natura delle caldere verte fin dall'inizio sulle stesse questioni sollevate dallo studio dei pennacchi, e cioè le condizioni termiche al di sotto della superficie e l'abbondanza di zolfo e di anidride solforosa. Possiamo chiamare le due interpretazioni più estreme delle caldere rispettivamente modello a zolfo e modello a silicati. Nel primo, proposto nella sua forma più semplice da Carl Sagan della Cornell University, i bacini sotterranei di zolfo fuso ipotizzati da Smith e collaboratori per spiegare i pennacchi di anidride solforosa di Io spiegano anche le colate vulcaniche visibili in superficie; le colate intorno alle caldere derivano da «lave» di zolfo liquido e non di basalto fuso o di altri silicati. Nel secondo modello, proposto per la prima volta da Harold Masursky e Michael H. Carr e collaboratori del Geological Survey, le caldere rappresentano una forma di vulcanismo silicatico molto simile a quello della Terra; l'unica differenza è che le colate di silicati di Io vengono colorate dallo zolfo.

Il modello a zolfo è basato sull'innegabile abbondanza su Io dello zolfo e dei suoi composti, sul fatto che è stata rilevata la presenza di anidride solforosa gassosa vicino a una delle eruzioni e sul fatto che probabilmente queste sostanze raggiungono il loro punto di fusione a piccole profondità nella crosta del satellite. Inoltre è stata avanzata l'ipotesi che la successione di mutamenti di colore e di viscosità dello zolfo al variare della sua temperatura possa spiegare il cambiamento di colore al crescere della distanza osservato intorno a molte delle caldere di Io. Un esempio tipico è Ra Patera. David C. Pieri del Jet Propulsion Laboratory e Sagan ipotizzano che le regioni scure al suo centro corrispondano alla fase nera ad alta temperatura dello zolfo fuso. Le colate lunghe che si irradiano come dita dalla caldera sarebbero flussi più freddi e meno viscosi delle fasi rossa e arancione.

Una difficoltà di questa ipotesi è che il comportamento delle varie fasi dello zolfo nelle condizioni di Io non è noto con sicurezza. Senza dubbio in laboratorio i colori delle fasi ad alta temperatura si possono conservare raffreddando rapidamente il liquido, ma le fasi così trattate sono solo metastabili e di solito si trasformano in zolfo solido normale dopo qualche tempo. Inoltre è probabile che su Io le fasi dello zolfo contengano numerose impurezze che, altrettanto probabilmente, causano ulteriori complicazioni.

L'argomentazione a favore del vulcanismo silicatico si basa in parte sui rilevamenti topografici di Io. Si tratta di misurazioni difficili perché le variazioni spaziali di luminosità nelle immagini riprese dai Voyager dovute a differenze nella composizione della superficie sono di solito molto più intense di quelle dovute al rilievo topografico. Ciononostante sono stati compiuti alcuni tentativi di misurazione su immagini ad alta risoluzione riprese in condizioni di illuminazione favorevole durante il periodo in cui *Voyager 1* si trovava alla distanza minima da Io. I risultati danno motivo di ritenere che Io abbia una topografia molto più varia degli altri satelliti galileiani, con montagne la cui altezza varia da 5000 a 10 000 metri e almeno qualche caldera profonda 2000-3000 metri. La conservazione di un rilievo così pronunciato sarebbe estremamente difficile in un corpo nella cui crosta si trovassero spessi depositi di zolfo e di anidride solforosa caldi, e forse liquidi. Sembrerebbe necessaria una crosta più resistente, con una struttura fatta perlomeno di silicati. La estesa distribuzione delle caldere sulla superficie di Io parrebbe, quindi, escludere l'esistenza di grandi «oceani» di zolfo sotto la sua superficie.

Questa sezione trasversale di Io rappresenta un tentativo di spiegare la sua attività vulcanica. In questo schema il mantello è composto di silicati fusi o parzialmente fusi (*in arancione*); la crosta è di silicati solidi (*in marrone*), ricoperta da strati di detriti silicatici ricchi di zolfo. In superficie lo zolfo è freddo e solido (*in giallo*), poi c'è una zona sottile in cui lo zolfo è fuso (*in rosso*) e infine una zona più spessa di zolfo molto caldo di consistenza simile alla pece (*in marrone giallastro*). A sinistra è raffigurato un pennacchio tipo Prometeo: lo zolfo fuso entra in contatto con l'anidride solforosa liquida (*in blu*) che bolle risalendo un condotto e forma in superficie un pennacchio ricco di neve di anidride solforosa. A destra è raffigurato un pennacchio analogo a Pele. Qui un'intrusione di silicati riscalda lo zolfo portandolo in una fase nera e catramosa (*in nero*), che alla fine evapora; il pennacchio quindi erutta vapori di zolfo. In alto a sinistra si vede un lago di lava come il lago Loki e, a destra in alto, una caldera, il cui pavimento è formato dalla crosta di silicati ricoperta da un sottile strato di zolfo.

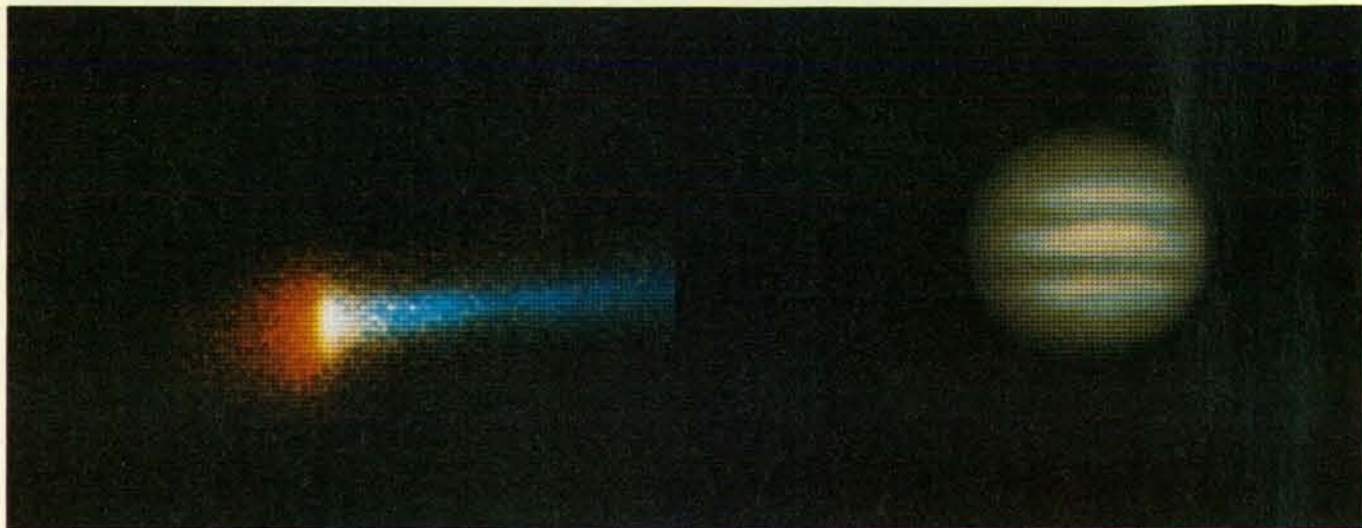
La termodinamica dell'anidride solforosa fa pensare che su Io sia possibile una vasta gamma di eruzioni a geyser. Sull'asse verticale è riportata la temperatura dell'anidride solforosa, su quello orizzontale l'entropia. (Le pressioni sono indicate dalle isobare in bar.) In un grafico di questo tipo la posizione orizzontale di una certa quantità di anidride solforosa rappresenta il rapporto tra le fasi (per esempio tra la fase liquida e quella gassosa) mentre una linea verticale rappresenta l'evoluzione di una eruzione di anidride solforosa che non compie lavoro (per esempio nel caso che incontri un ostacolo) e quindi non acquisisce né cede calore. Nel diagramma sono indicate due eruzioni di questo tipo. La prima (*a*) inizia quando la temperatura dell'anidride solforosa liquida sale a 393 kelvin, ossia raggiunge la temperatura dello zolfo fuso nelle regioni più profonde della crosta di Io. L'anidride solforosa, di conseguenza, bolle e sale verso la superficie. La seconda eruzione (*b*) inizia con anidride solforosa gassosa. In entrambi i casi l'anidride solforosa esplode in una nube di ghiaccio e gas nello spazio quasi vuoto al di sopra della superficie di Io. Questa analisi termodinamica è stata compiuta da Susan W. Kieffer dello US Geological Survey.

L'altra argomentazione principale avanzata dai sostenitori del vulcanismo silicatico è che le strutture visibili su Io (le caldere, le lunghe colate a forma di dita, i coni vulcanici con crateri centrali e così via) sono più o meno le stesse che i geologi sono abituati a vedere sulla Terra e su Marte. Sembra ben poco plausibile che lo zolfo, un materiale vulcanico inconsueto con una relazione temperatura-viscosità fuori dall'ordinario, possa riprodurre così bene le strutture generate dalle lave basaltiche. Sfortunatamente i geologi hanno solo una scarsa esperienza diretta dello zolfo come fluido vulcanico. Sulla Terra sono rare le colate di zolfo; inoltre avvengono solo quando il vulcanismo silicatico fonde depositi di zolfo.

Che cosa succede, allora, su Io? Nella contesa tra zolfo e silicati entrambe le fazioni sono oggi disposte ad ammettere che la forma di attività vulcanica sostenuta dalla fazione opposta avvenga su Io almeno in qualche caso. Gerald G. Schaber del Geological Survey, ad esempio, osserva che alcune caldere si sono formate in regioni montuose del satellite. Le montagne sono presumibilmente composte da silicati, quindi anche le caldere devono esserlo. Inoltre tutte e due le parti sono concordi nell'ammettere che l'attività vulcanica alla superficie di Io sia alimentata, indipendentemente dalla sua natura, da silicati fusi presenti in profondità nella crosta di Io; il problema è se uno dei due tipi di fluido vulcanico, lo zolfo o i silicati, risulti dominante in superficie.

Rimane da descrivere ancora una struttura vulcanica. A sud dei pennacchi Loki esiste una formazione isolata scura, probabilmente un grande lago di lava come quelli che di solito riempiono le caldere attive sulla Terra durante le eruzioni. All'epoca del passaggio di *Voyager 1* questa struttura era il più grande «punto caldo» di Io, con una temperatura di circa 300 kelvin, mentre la temperatura di fondo locale era di soli 130 kelvin. Inoltre le immagini ad alta risoluzione hanno rivelato che all'interno della struttura esisteva una «zattera» di materiale chiaro, apparentemente solcata da crepe e circondata da frammenti più piccoli dello stesso materiale che sembravano essersi staccati dai bordi. È come se la crosta in via di raffreddamento della struttura fosse stata frantumata dai moti convettivi o dall'aggiunta di altro materiale eruttivo. La struttura è molto più estesa dei laghi di lava delle caldere hawaiiane, e anzi, con i suoi 250 chilometri circa di lunghezza, potrebbe contenere l'intero arcipelago. È piena di silicati fusi o di zolfo fuso ormai raffreddato? Non si sa.

La prova più chiara del fatto che la superficie di Io è giovane dal punto di vista geologico è l'assenza totale in tutte le immagini riprese dai Voyager di crateri da impatto. Altrove i geologi planetari avevano trovato queste tracce di bombardamenti antichi e attuali dovunque erano riusciti a cercarle, anche su Marte e sulla Terra, e le superfici butterate di Ganimede e Callisto indicano che anche i satelliti di Giove hanno avuto una storia di impatti simile a quella dei pianeti interni. Inoltre i calcoli di Shoemaker suggeriscono che il flusso attuale di comete e asteroidi attraverso il sistema gioviano dovrebbe continuare a produrre su Io crateri di grandi dimensioni a un tasso analogo, entro un ordine di grandezza, a quello valido per la Luna.

Sulla base dei risultati di Shoemaker abbiamo calcolato quanta attività vulcanica è necessaria per seppellire i crateri di Io. Siamo arrivati alla conclusione che in media su Io deve spargersi un millimetro di spessore di detriti vulcanici all'anno. Questo implica che il satellite è un corpo con un'attività vulcanica più che notevole. Quantitativamente, vuol dire che Io erutta diverse migliaia di tonnellate di

Un toro di zolfo ionizzato centrato sull'orbita di Io circonda Giove. Il toro deriva dall'attività vulcanica del satellite, che si ritiene emetta nello spazio circostante circa una tonnellata di atomi di zolfo e di ossigeno al secondo. L'immagine di questa porzione di toro è stata ottenuta ai Mount Wilson and Las Campanas Observatories da John T. Trauger del Cal Tech. I falsi colori indicano le proporzioni relative degli atomi di zolfo ionizzati una volta (*in blu*) e due volte (*in rosso*). Il margine apparente, a destra, segna il limite dei dati di Trauger.

materiale al secondo, cioè ogni mese circa il quantitativo totale di materiale espulso dall'eruzione del Mount St. Helens nel maggio 1980. Un ammontare così elevato ha conseguenze notevoli: tra l'altro fa pensare che almeno gli strati superiori del mantello e della crosta di Io siano stati riciclati molte volte nel corso della storia del satellite. Questo fenomeno concorda con l'idea che Io abbia perduto completamente, a causa del riscaldamento mareale, tutte le sostanze volatili che conteneva inizialmente.

I particolari di questo riciclaggio non sono ancora ben noti. Non è chiaro, per esempio, se nel processo di ricopertura della superficie siano più importanti i pennacchi o le colate laviche. Le valutazioni del quantitativo totale di materiale eruttato nei pennacchi di Io attribuiscono loro un tasso di ricostituzione della superficie compreso tra qualche decimillesimo di millimetro all'anno e un decimo di millimetro o più. Questo fa pensare che le colate siano per lo meno su un piano di parità con i pennacchi. Naturalmente, parlando di queste stime, non bisogna dimenticare che Io è estremamente dinamico e che quindi i due passaggi dei Voyager ce ne hanno offerto, per così dire, solo due istantanee.

Sono visibili dalla Terra segni dell'attività vulcanica di Io? Stranamente la risposta è affermativa. In effetti esistevano prove dell'esistenza di vulcani su Io in una serie di dati enigmatici raccolti diversi anni prima dei passaggi dei Voyager. Le osservazioni telescopiche di Io nell'infrarosso compiute alla fine degli anni sessanta e nei primi anni settanta rivelarono due fatti strani a proposito di questo corpo. Innanzitutto, quando Io viene eclissato da Giove la sua temperatura cade molto rapidamente, come sarebbe prevedibile per un corpo privo o quasi di atmosfera con una superficie costituita da uno strato di polvere fine isolante. Su Io, però, contrariamente a quanto accade sulla Luna, la temperatura minima raggiunta dal satellite quando è profondamente immerso nell'ombra di Giove è troppo elevata per essere compatibile con l'esistenza di uno strato isolante omogeneo. In secondo luogo la temperatura di Io che si deduce dalla sua luminosità infrarossa non è la stessa a tutte le lunghezze d'onda, ma è decisamente più elevata alle lunghezze d'onda più corte.

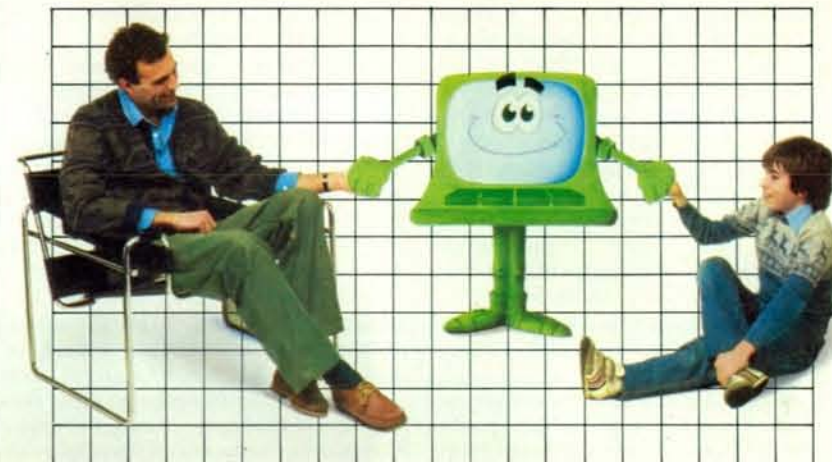
Ciascuno di questi problemi, preso isolatamente, poteva essere affrontato, anche se con sistemi *ad hoc*. Per spiegare i dati delle eclissi vennero elaborati modelli a due strati del satellite e per spiegare le anomalie termiche fu ipotizzata una variazione dell'emissività di Io in funzione delle lunghezze d'onda. Dopo i passaggi dei Voyager e la scoperta di vulcani su Io tre ricercatori del Jet Propulsion Laboratory, Dennis L. Matson, Gary A. Ransford e uno di noi (Johnson), tornarono a esaminare quei dati, con la speranza di risolvere i vecchi problemi alla luce delle nuove informazioni, ed elaborarono un semplice modello in cui la superficie isolante di Io è cosparsa di punti caldi di temperatura data. Si scoprì che i problemi scomparirebbero se circa l'1 per cento della superficie di Io fosse ricoperto di punti caldi aventi una temperatura analoga a quella misurata dalle sonde Voyager.

Il modello ha fornito anche un valore per la quantità di energia irradiata dai punti caldi: mediandola sulla superficie di Io arrivava a due watt circa per metro quadrato. Una nuova analisi dei dati dei Voyager e una nuova serie di osservazioni nell'infrarosso eseguite dal Mauna Kea forniscono due valutazioni indipendenti dello stesso parametro; entrambe sono in accordo con il modello e con i dati precedenti. Oggi la valutazione migliore del flusso di calore medio proveniente da Io, tenendo conto di tutte queste fonti, è di 1,5 watt per metro quadrato più o meno 0,5 watt. La Terra, a paragone, irradia 0,08 watt per metro quadrato e la Luna, che ha quasi la stessa massa e le stesse dimensioni di Io, solo 0,03 watt. La radiazione di Io si distingue per essere completamente anomala: anche escludendo il riscaldamento mareale sarebbe chiaro

Il raffreddamento di Io durante le eclissi ha rappresentato a lungo un vero enigma per gli osservatori che misuravano il flusso infrarosso del satellite mentre si trovava nell'ombra di Giove. I ricercatori prevedevano che il flusso sarebbe sceso rapidamente verso lo zero all'inizio di ogni eclisse, per ritornare al valore normale quando la luce solare avesse riscaldato nuovamente la superficie del satellite (*curva in nero*). Invece il flusso rimaneva al di sopra dello zero (*curva in colore*). Questo succede perché i «punti caldi» vulcanici di Io emettono un flusso di calore più o meno costante per tutta la durata dell'eclisse. Quindi l'attività vulcanica di Io può essere seguita anche dalla Terra.

Il BASIC in meno di 6 mesi

"Datemi una leva e solleverò il mondo"!
Noi del Gruppo Editoriale Jackson siamo più concreti:
"Dateci meno di 6 mesi di tempo e vi insegneremo il BASIC"! In modo facile e completo con

ABC Personal Computer

ABC PERSONAL COMPUTER il nuovissimo corso programmato in cui troverete tutto quello che dovete sapere sul mondo dei Personal e sul loro linguaggio fondamentale: il BASIC!

ABC PERSONAL COMPUTER il facilissimo corso tenuto da esperti nel quale la grammatica va d'accordo con la pratica: c'è l'hardware e il software, i codici e i loro comandi, i programmi e come si impostano; si parla di differenze, pregi e difetti dei Personal in commercio... In più: consigli pratici, storia dei calcolatori, centinaia di esempi scritti, schemi, grafici e fotografie.

ABC PERSONAL COMPUTER in tutto 24 fascicoli settimanali da rilegare in 4 preziosi e ricchi volumi:
- 2 volumi di Lezioni per 608 pagine complessive
- 1 volume di Computer Test di 108 pagine
- 1 Dizionario di Informatica di 200 pagine

ABC Personal Computer una pubblicazione del Gruppo Editoriale Jackson

In edicola il 1° fascicolo con il piano dell'opera e un Poster in regalo!

GRUPPO EDITORIALE JACKSON **"noi l'informatica la conosciamo davvero!"**

nuovidea

che su Io agisce una fonte di energia diversa dalla radioattività interna.

Va detto che il flusso di calore di Io differisce da quello della Terra e della Luna non solo quantitativamente, ma anche qualitativamente: sulla Terra e sulla Luna quasi tutto il calore interno in eccesso generato dalla radioattività arriva in superficie per conduzione attraverso la crosta. (Le eruzioni vulcaniche sulla Terra contribuiscono al totale solo in piccola parte.) Dalla superficie il calore viene irradiato nello spazio senza elevare in maniera apprezzabile la temperatura del corpo. (La superficie, infatti, viene riscaldata soprattutto dall'energia che riceve dal Sole.) Questo significa che l'eccesso di energia non si può misurare a distanza come emissione infrarossa, ma solo misurando con precisione il gradiente termico nella parte superiore della crosta. Su Io le cose vanno diversamente: il calore interno in eccesso arriva in superficie soprattutto per convezione, trasportato dai fluidi ad alta temperatura che salgono dai punti caldi. Da questi punti il calore viene reirradiato nello spazio a temperature molto superiori a quella superficiale media del satellite, e risulta facile da misurare. L'emissione di energia dalla superficie di Io è compresa tra 10^{13} e 10^{14} watt.

Fino a oggi, oltre alla radioattività interna, sono stati proposti come possibili sorgenti di questa energia due processi. Uno è il riscaldamento mareale; l'altro è il riscaldamento che Io potrebbe subire per effetto della sua resistenza alle correnti elettriche indotte nel suo interno dall'interazione con la magnetosfera di Giove. In effetti, Thomas Gold della Cornell University avanza l'ipotesi che i pennacchi di Io derivino da un'interazione di questo tipo che ha luogo quando la materia ionizzata sale lungo le linee di campo della corrente che attraversano la superficie di Io come grandi fulmini persistenti. La sua proposta si basa da un lato sul fatto che è difficile spingere i gas vulcanici a velocità superiori a quella del suono e, dall'altro, sui dati dei magnetometri dei Voyager dai quali risulta che nelle vicinanze di Io passi una corrente di circa un milione di ampere lungo le linee del campo magnetico. (Un milione di ampere è anche la corrente che ci si aspetterebbe dal calcolo teorico del flusso di ioni nella magnetosfera di Giove.)

Il fatto che Io emetta 10^{14} watt si può sfruttare per verificare l'eventuale importanza degli effetti elettromagnetici nel suo bilancio energetico complessivo. Supponiamo che tutta la corrente che si calcola incontri Io venga convertita in calore nel suo interno per effetto della resistenza. L'entità netta del riscaldamento sarebbe compresa tra 10^{11} e 10^{12} watt, non più di un centesimo di quella necessaria per spiegare l'emissione del satellite. Gli effetti elettromagnetici potrebbero essere comunque importanti per fenomeni come i pennacchi eruttivi. In questo caso, però, il «fulmine» che genera ciascun pennacchio produrrebbe su Io un punto caldo con una temperatura effettiva dell'ordine dei 100 000 kelvin. La ricerca di punti caldi simili sulla faccia in ombra di Io non ha avuto successo.

Nel frattempo è sorto un problema a proposito dell'ipotesi del riscaldamento mareale. L'interazione mareale tra Giove e Io, infatti, implica lo sviluppo di protuberanze mareali sia sul satellite sia sul pianeta; quella sul pianeta applica a Io una coppia gravitazionale che lo fa accelerare e spostare su orbite sempre più alte. (Un'interazione analoga è responsabile del lento, ma continuo allontanamento della Luna dalla Terra.) Le migliori valutazioni oggi disponibili dell'intensità dell'interazione stabiliscono un limite superiore per l'apporto medio di energia al satellite e questo limite è circa due volte inferiore al flusso di calore misurato proveniente da Io.

Il problema è ancora irrisolto. Forse il flusso termico e l'attività vulcanica di Io sono estremamente variabili nel tempo e i Voyager sono arrivati in un periodo in cui l'attività era eccezionalmente rilevante. Forse alcuni dei parametri e delle ipotesi utilizzate per il calcolo dell'evoluzione dell'orbita e del riscaldamento mareale sono sbagliati. O forse i punti caldi non sono distribuiti uniformemente sulla superficie del satellite. I dati raccolti ultimamente all'Infrared Telescope Facility della NASA ad Hawaii da alcuni ricercatori del Jet Propulsion Laboratory tra cui uno di noi (Johnson), fanno pensare che la regione intorno a Loki contribuisca forse con una percentuale notevole al flusso termico totale di Io. Se è così, e se quindi Loki e i suoi dintorni sono stati erroneamente considerati rappresentativi di tutto il satellite, l'energia totale emessa da Io è forse più vicina alle previsioni teoriche.

Non sembra che esistano prove di variazioni notevoli nel flusso di calore di Io: i dati registrati nei primi anni settanta durante eclissi sono in accordo con le misurazioni dei Voyager e con i dati registrati nel corso di eclissi negli anni ottanta. Un tipo di radiazione infrarossa mostra in realtà variazioni impressionanti sul breve termine: l'emissione nella regione dello spettro vicina alla lunghezza d'onda di cinque micrometri rivela impulsi occasionali legati probabilmente a processi che avvengono su Io più o meno quotidianamente.

Il primo di questi impulsi è stato rilevato nel 1978 da Fred C. Witteborn dell'Ames Research Center e collaboratori a bordo dell'aeroplano di alta quota, appositamente attrezzato, chiamato Kuiper Airborne Observatory. La loro breve osservazione di un flusso a cinque micrometri, di intensità più che doppia di quella che si poteva attribuire alla luce solare riflessa da Io, causò notevole sorpresa.

Questo flusso fece pensare che una piccola area pari a circa un decimillesimo della superficie di Io avesse una temperatura di circa 600 kelvin, ma a quell'epoca l'esistenza di un'attività vulcanica sembrava quasi inconcepibile. Dopo il passaggio di *Voyager 1* l'importanza dell'osservazione di Witteborn apparve immediatamente chiara: lo spettrometro infrarosso di *Voyager 1* aveva individuato piccole regioni di Io con temperature analoghe. William Sinton, lavorando al Mauna Kea, avviò una ricerca sistematica di altri «eventi a cinque micrometri». Uno dei suoi primi successi fu la scoperta dell'evento nell'arco di tempo compreso tra i due passaggi dei Voyager, quello che si ritiene sia associato ai mutamenti intorno alla caldera Surt.

Da allora Sinton ha continuato ad accumulare diversi anni di osservazioni, trovando pochi eventi in grado di rivaleggiare con i primi; i suoi dati però rivelano una variazione continua del flusso proveniente dai più caldi tra i «punti caldi» di Io. La quantità totale di energia irradiata da queste piccole regioni ad alta temperatura è molto inferiore a quella emessa da zone più estese a temperature inferiori: per questo motivo gli eventi a cinque micrometri non contribuiscono in maniera significativa al bilancio energetico complessivo del satellite. Ciononostante, questi fenomeni aiutano a comprendere gli eventi di grande violenza che presumibilmente generano grandi pennacchi di tipo Pele e, in futuro, faciliteranno la ricerca di indizi dell'esistenza di lave silicatiche. Più in generale le osservazioni di Io compiute dalla Terra, compresi gli studi delle eclissi e le misurazioni nell'infrarosso, potenzialmente sono tutte metodi per seguire l'evoluzione dell'attività vulcanica di Io e costituiranno un importante raccordo tra i dati dei Voyager e quelli che saranno raccolti dalla prossima sonda che raggiungerà il sistema gioviano.

Il prossimo veicolo spaziale sarà *Galileo*, il cui lancio è previsto per il 1986, con arrivo nelle vicinanze di Giove nell'agosto 1988. *Galileo* sarà composto di una sonda per penetrare nell'atmosfera di Giove e di un modulo orbitante in grado di fornire una ventina di mesi di osservazioni dettagliate del pianeta, della sua magnetosfera e dei suoi satelliti. Nel suo viaggio verso Giove il veicolo passerà a meno di 1000 chilometri da Io, cioè 20 volte più vicino dei Voyager. Gli strumenti a bordo di *Galileo* costruiranno una mappa del satellite con immagini di risoluzione paragonabile a quelle della Terra ottenute dai Landsat. Verrà esaminata l'interazione tra la magnetosfera di Giove e l'atmosfera rarefatta di Io, mentre i rilevamenti radio della traiettoria del veicolo porranno nuovi vincoli ai modelli dell'interno del satellite.

Dopo la fase di massimo avvicinamento, *Galileo* dovrà tenersi lontano dai dintorni di Io, altrimenti verrebbe danneggiato gravemente da un'esposizione prolungata alle radiazioni. Ciononostante per la maggior parte dei venti mesi della sua missione la sonda rimarrà marcatamente all'interno della distanza da cui i Voyager hanno individuato e osservato i vulcani di Io. E per una decina di volte o più *Galileo* sarà in grado di esplorare il satellite in cerca di mutamenti dell'attività vulcanica, di farne mappe termiche che indichino le posizioni dei punti caldi e di cercare tracce del passaggio di ioni da Io alla magnetosfera di Giove.

SARA' COPIATA. INEVITABILE.

Ci sono auto che lasciano il segno. Sono i modelli che imprimono una svolta decisiva nella evoluzione progettuale dell'automobile. Sono i modelli che dettano i canoni stilistici e funzionali cui si conformano inevitabilmente altre vetture.

Sono modelli unici, inconfondibili, irripetibili. Esempio da imitare, linea da seguire. La linea di Sierra. Espressione della ricerca più avanzata applicata alla produzione di serie, Sierra traduce nella straordinaria efficienza aerodinamica un nuovo concetto di guida.

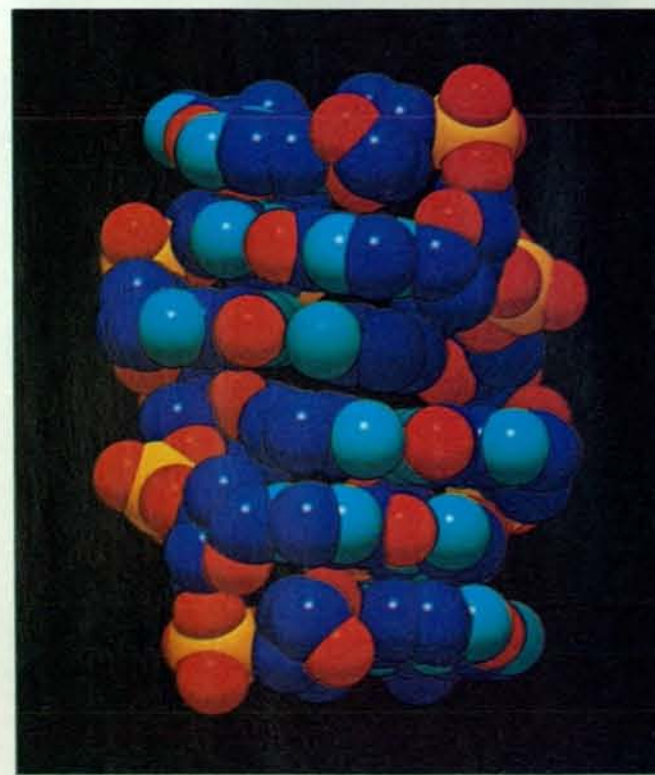
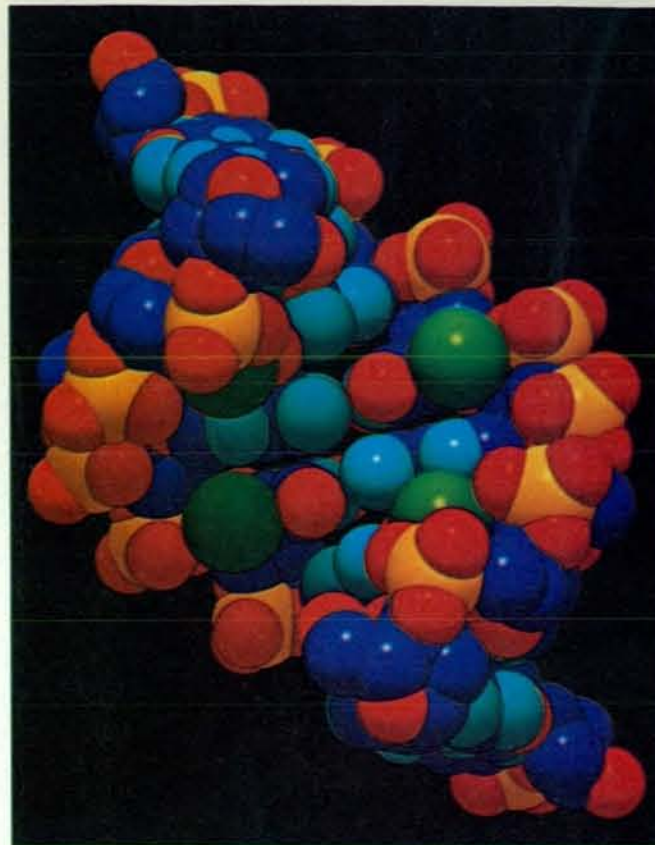
L'eliminazione di qualsiasi sporgenza anche attraverso innovativi metodi di incollatura del parabrezza, i paraurti integrati nel corpo vettura, un equipaggiamento esclusivo studiato in funzione ergonomica, un sistema di sospensioni che esalta l'elasticità e la progressione dei propulsori a benzina o diesel: in ogni particolare Sierra interpreta l'impegno Ford nell'elaborazione di nuove tecnologie automobilistiche.

Una sofisticata metodologia costruttiva che trova costante applicazione su tutti i modelli dell'intera gamma Ford.

Perchè ogni automobile Ford possa rispondere a sempre più avanzate e specifiche esigenze automobilistiche.

Perchè l'automobile continui ad essere un veicolo del progresso.

Tecnologia e temperamento.

In questi modelli molecolari compatti, che sono immagini al calcolatore generate da Nelson Max del Lawrence Livermore National Laboratory, sono rappresentati tre tipi di doppia elica. Si tratta di tre corte molecole a doppia elica, le cui strutture sono state risolte mediante analisi diffrattometrica con i raggi X, effettuata su cristalli singoli. Gli atomi di carbonio sono in blu intenso, quelli di azoto in blu chiaro, quelli di ossigeno in rosso e quelli di fosforo in giallo. Gli atomi di bromo presenti in alcune basi sono verdi. Il DNA *B* (*a sinistra*) viene rappresentato da una molecola a 12 basi, la cui struttura è stata risolta dall'autore e dai suoi collaboratori. L'impalcatura di sostegno fatta di fosfati dà una chiara idea dell'avvolgimento (o avvitamento) destrorso. La molecola a 8 basi del DNA *A*, in alto a destra, è stata risolta da Olga Kennard, Zippora Shakked e da M. A. Viswamitra. Anch'essa ha un avvolgimento destrorso. L'immagine, che mette in mostra il profondissimo solco principale, evidenzia in che modo le coppie di basi sono inclinate rispetto all'asse verticale dell'elica. La molecola del DNA *Z*, in basso a destra, ha una struttura che è stata risolta da Andrew H.-J. Wang, da Alexander Rich e dai loro collaboratori. Come indicano i gruppi fosfato, a destra in alto e a sinistra in basso dell'immagine, l'elica del DNA *Z* ha un avvitamento sinistrorso. L'impalcatura di sostegno ha un andamento a zig-zag e le coppie di basi sono impaccate due a due invece che singolarmente, il che conferisce all'elica una struttura alternata.

Come viene letta l'elica del DNA

L'analisi ai raggi X dei cristalli di tre tipi di DNA a doppia elica permette di concludere che l'informazione contenuta nella sequenza delle basi può essere immagazzinata nella struttura locale dell'elica

di Richard E. Dickerson

Nella doppia elica del DNA sono contenuti due tipi di informazione genetica, i quali sono codificati e interpretati in modi molto diversi. Il messaggio genetico stesso, cioè l'informazione che specifica la struttura delle proteine, viene scritto nel ben noto codice genetico a triplette. Questo codice è lineare e la sua interpretazione è estrinseca. La sequenza in cui successive triplette dei gruppi chimici che vengono chiamati basi sono schierate lungo un singolo filamento dell'elica codifica per la sequenza in cui successivi amminoacidi sono uniti a formare una catena proteica. Non sembra esservi tra una particolare tripletta di basi e l'amminoacido che essa specifica alcuna relazione strutturale innata, ma piuttosto il trasferimento dell'informazione viene mediato indirettamente da quel complesso meccanismo che è estrinseco al DNA e che è costituito dall'RNA messaggero, dai ribosomi, dall'RNA di trasporto e da una serie di appositi enzimi.

Il DNA contiene in codice non solo il messaggio genetico, ma anche le istruzioni per la sua espressione selettiva. Il poco che si conosce su questo tipo di informazione proviene da studi sul controllo genetico nei batteri. Sul cromosoma batterico blocchi di geni sono disattivati dal legame di una proteina, che funge da repressore, a una regione del DNA (l'operatore), che possiede una speciale sequenza di basi che quel repressore riconosce; i geni diventano attivi quando varie piccole molecole si legano al repressore e lo staccano dall'operatore. Benché molto rimanga ancora da imparare, la maggior parte degli studiosi di biologia molecolare pensa che meccanismi analoghi siano alla base del controllo genetico in forme di vita superiori rispetto ai batteri. Un repressore batterico riconosce l'operatore direttamente, ed è probabile che formi legami a idrogeno tra gli atomi di azoto e di ossigeno presenti nei suoi amminoacidi e sui margini delle coppie complementari di basi della doppia elica del DNA. Questa si adatta comodamente a una particolare conformazione della proteina. Pertanto, l'informazione che specifica il controllo genetico è tridimensionale e la sua interpretazione è intrinseca; essa dipende dalle proprietà strutturali implicite della proteina e dell'elica.

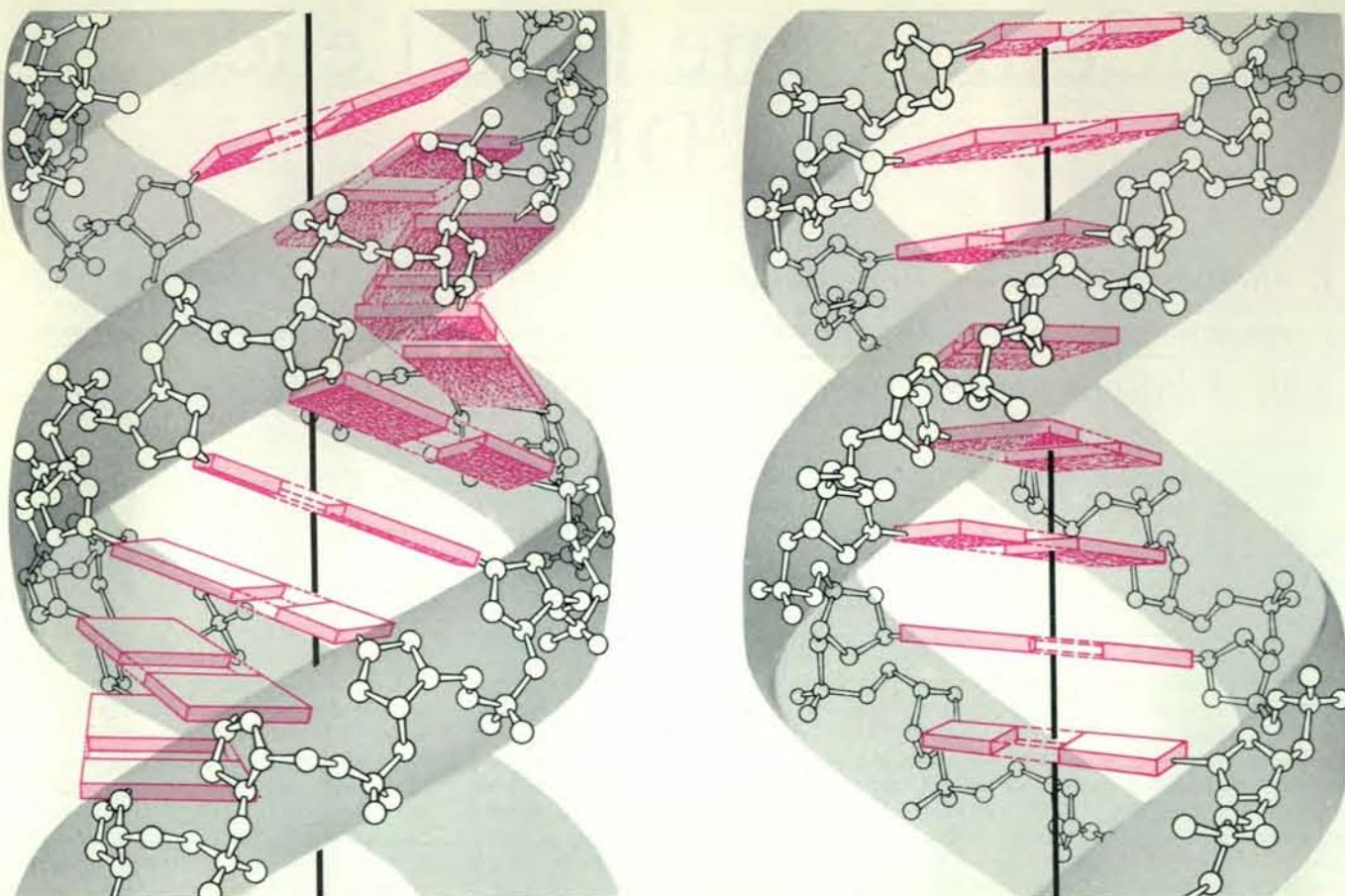
Nel processo di riconoscimento, il DNA svolge solo un ruolo passivo, con le catene laterali degli amminoacidi che si inseriscono nei solchi di una doppia elica statica, oppure la stessa sequenza di basi modifica la struttura dell'elica in un modo che contribuisca al riconoscimento di siti specifici da parte delle proteine di controllo? Quattro anni fa soltanto la domanda sarebbe stata oziosa, in quanto, in base al tipo di dati disponibili allora, non era possibile darle una risposta.

James D. Watson e Francis Crick proposero la loro struttura a doppia elica per il DNA nel 1953, sulla base di fotografie di diffrazione dei raggi X, ottenute da Rosalind Franklin e Maurice Wilkins. Tali fotografie si riferivano a fibre di DNA. Un'ulteriore analisi dei diffrattogrammi permise di concludere che le fibre del DNA possono esistere in due forme: come DNA *B* in condizioni di umidità elevata e come DNA *A* quando l'umidità è più bassa. Per ambedue queste forme furono sviluppati modelli molecolari. Tuttavia, la quantità di informazione strutturale che si poteva ricavare dai diffrattogrammi era intrinsecamente limitata dal disordine dei singoli filamenti di DNA tutt'attorno e lungo l'asse della fibra. Al massimo si poteva stabilire una struttura ad elica globale media. Una qualsiasi variazione locale nella struttura, indotta ad esempio da una certa sequenza di basi, non poteva essere individuata.

Oggi, invece, variazioni di questo tipo possono essere individuate e misurate con precisione. Metodi avanzati per la sintesi organica di corte molecole di DNA, con una sequenza desiderata qualsiasi, hanno permesso per la prima volta di produrre oligomeri di DNA (corti filamenti di DNA) con un numero di basi da quattro a 24 in quantità sufficiente e sufficientemente pure da poter essere cristallizzate e studiate con i metodi convenzionali di diffrazione dei raggi X da parte di cristalli singoli. Questo articolo vuol essere appunto una finestra aperta su alcuni degli interessantissimi nuovi risultati che cominciano a trasparire dall'analisi strutturale di corte molecole di DNA a doppia elica.

La diffrazione nelle fibre

L'esame delle immagini di diffrazione dei raggi X ottenute con fibre stirate e con sottili pellicole di DNA naturale ha rivelato la struttura di base dei due tipi fondamentali di doppia elica: la configurazione *B*, che è stabile a una umidità relativa di circa il 92 per cento, e la configurazione *A*, che la maggior parte delle sequenze di basi assume quando l'umidità scende a circa il 75 per cento. Ambedue hanno la forma di una scala a pioli flessibile, avvolta ad elica attorno a un asse centrale. I due montanti della scala sono catene costituite da anelli dello zucchero desossiribosio alternati a gruppi fosfato. I pioli sono, invece, coppie di basi puriniche e pirimidiniche. Vi sono due tipi di basi puriniche a doppio anello, l'adenina (*A*) e la guanina (*G*), così come vi sono due tipi di pirimidine ad anello singolo, la timina (*T*) e la citosina (*C*). Queste coppie di basi sono tenute unite da legami a idrogeno. L'adenina si appaia di norma con la timina mediante due legami a idrogeno e la guanina si appaia con la citosina mediante tre legami a idrogeno.

I due punti di attacco di una coppia di basi agli anelli del desossiribosio non sono situati l'uno diametralmente opposto all'altro sulla linea che interseca la coppia di basi (*si veda l'illustrazione in basso a pagina 69*) e questo fatto è importante per la geometria della doppia elica del DNA. Il margine delle coppie di basi dalla cui parte l'angolo compreso tra gli attacchi è inferiore a 180 gradi si chiama margine del solco minore, mentre quello che forma un

Le analisi su fibre di DNA di Struther Arnott e collaboratori, precedenti agli studi sui cristalli singoli di cui si parla in questo articolo, sono state alla base di questi disegni del DNA *A* (*a sinistra*) e *B* (*a destra*), realizzati da Irving Geis. Le impalcature zucchero-fosfato della doppia elica sono rappresentate come nastri e le coppie di basi simili a pioli, che le connettono, come tavole. Nel DNA *A* le coppie di basi sono inclinate e sospinte lontano dall'asse della doppia elica. Nel DNA *B*, invece, giacciono a cavallo dell'asse dell'elica e sono perpendicolari ad esso.

angolo superiore a 180 gradi è detto margine del solco maggiore. Quando le coppie di basi sono sovrapposte l'una all'altra in un'elica, le impalcature di sostegno costituite dai gruppi fosfato formano le due pareti di un solco maggiore e di un solco minore, che si avvolgono attorno all'elica, mentre i margini delle basi formano il pavimento dei solchi.

Il pavimento del solco maggiore è lastricato di atomi di azoto e di ossigeno che possono formare legami a idrogeno con le catene laterali degli amminoacidi di una proteina e pertanto possono avere un ruolo significativo nella codificazione intrinseca. La disposizione di questi gruppi che formano legami a idrogeno è diversa per i due tipi di coppie di basi. Passando da una purina a una pirimidina, una coppia *A–T* offre un atomo di azoto (accettore di idrogeno), un gruppo NH_2 (donatore) e un atomo di ossigeno (altro accettore). Per contro, la coppia *G–C* offre gli stessi gruppi in un ordine diverso: dapprima un azoto (accettore), quindi un ossigeno (accettore) e infine un NH_2 (donatore). Dato che ogni coppia di basi (*A–T* e *G–C*) può anche essere ribaltata di 180 gradi (e diventare *T–A* e *C–G*), a ogni giro dell'elica possono venir mostrate al repressore, o a un'altra proteina di controllo, quattro differenti combinazioni. Il pavimento del solco maggiore porta pertanto un messaggio, la sequenza di basi del DNA, in una forma che può essere letta soltanto da altre grosse molecole.

Il solco minore contiene meno informazione: lo schema dei legami a idrogeno in coppie di basi *A–T* è semplicemente accettore-accettore, indipendentemente dal senso in cui la coppia di basi è orientata, e la coppia *G–C* differisce soltanto per l'intrusione di un donatore NH_2 tra gli accettori. Ciò rende il solco minore un candidato meno probabile per la presentazione dell'informazione. Come risulterà chiaro più avanti, questo solco ha un'altra importante funzione nel DNA *B*.

Il DNA *A* e il DNA *B* differiscono principalmente nella posizione delle coppie di basi rispetto all'asse dell'elica e nell'inclinazione delle basi stesse, cioè nella pendenza in sezione della coppia di basi rispetto all'asse longitudinale del DNA. Nel DNA *B* quest'inclinazione è vicina a zero (cioè le coppie di basi risultano sovrapposte l'una all'altra quasi perpendicolarmente all'asse dell'elica) e l'asse passa attraverso il centro di ogni coppia di basi. Il solco minore è più stretto del maggiore a causa dell'attacco asimmetrico delle coppie di basi agli anelli di desossiribosio, ma i due tipi di solco hanno profondità simile: grosso modo la stessa distanza in profondità intercorre dalla superficie del cilindro che contiene l'elica fino al margine di una coppia di basi. Nel DNA *A*, d'altra parte, le coppie di basi sono inclinate, rispetto alla posizione perpendicolare, di 13-19 gradi. Inoltre, esse sono spostate verso l'esterno dell'elica, il cui asse giace nel solco maggiore, passando al di fuori delle basi. Pertanto il solco minore è poco profondo, è poco più di una semplice depressione che si avvolge a elica attorno alla superficie esterna del cilindro; il solco maggiore è molto profondo, estendendosi, per tutto il tragitto, dalla superficie oltre l'asse centrale e, in parte, anche verso il lato opposto.

Il DNA *B* ha una media di 10 coppie di basi per giro d'elica, con una spaziatura di 0,34 nanometri lungo l'asse dell'elica, da una coppia di basi alla successiva. L'elica *A* ha quasi 11 coppie di basi per giro e, dato il modo in cui sono sovrapposte l'una all'altra le coppie di basi inclinate, la distanza lungo l'asse è di soli 0,29 nanometri per coppia di basi. Questi valori sono delle medie, derivate dai dati sulle fibre; l'analisi effettuata su cristalli singoli ha rivelato alla fine grosse deviazioni locali. (Il modello originale di Watson e Crick aveva le 10 coppie di basi per giro e la sovrapposizione perpendicolare delle basi che sono caratteristiche del DNA *B*, ma le basi erano tenute lontane dal centro per lasciare una cavità centrale che era, in effetti, più tipica del DNA *A*. Anche la posizione degli anelli di desossiribosio nel modello originale era più vicina a quella della forma *A*.)

Questo fu pressapoco lo stato delle conoscenze sulla doppia elica del DNA nei 25 anni che seguirono alla scoperta di Watson e Crick. Molte domande rimanevano senza risposta e senza possibilità di essere risolte. Perché l'elica *B* è la forma dell'umidità elevata? Quale forma prevale nel DNA cromosomico e può una forma essere convertita in un'altra all'interno di un organismo vivente? Le proteine di riconoscimento, ad esempio i repressori, possono realmente leggere l'informazione dal fondo dei solchi quando si legano a sequenze specifiche di DNA?

Alla fine degli anni settanta, quando i progressi compiuti nella sintesi del DNA permisero per la prima volta di studiare cristalli singoli di corte molecole con una sequenza prescelta qualsiasi, sia il gruppo di Alexander Rich al Massachusetts Institute of Technology sia il nostro gruppo, che allora lavorava al California Institute of Technology, incominciarono a sintetizzare indipendentemente l'uno dall'altro dei copolimeri *C–G*: brevi molecole composte esclusivamente da molecole di citosina e da molecole di guanina che si alternavano. C'era un motivo per la scelta di questa sequenza. Esperimenti effettuati con soluzioni avevano suggerito che il copolimero *C–G*, in condizioni di elevata concentrazione di sali o di alcool, va incontro a un certo tipo di transizione strutturale e, pertanto, era il candidato più probabile per lo studio delle transizioni tra i vari tipi di elica.

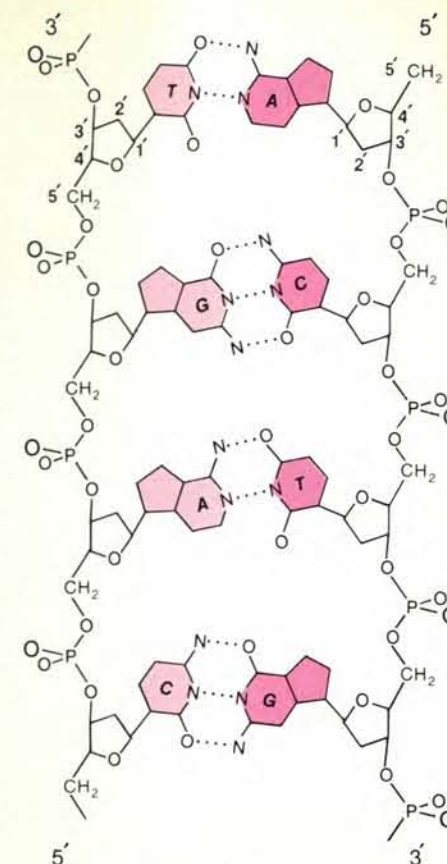
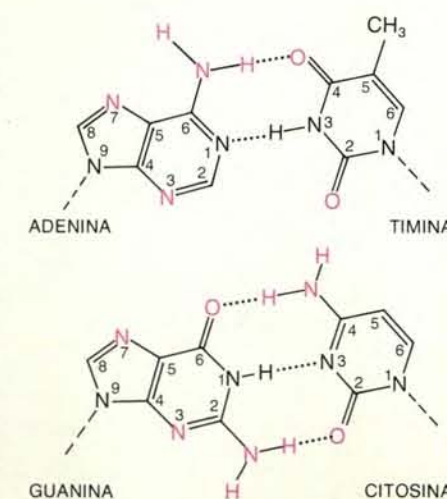
Gli studi con cristalli singoli

Andrew H.-J. Wang del Massachusetts Institute of Technology risolse la struttura dell'esamero *CGCGCG* verso la metà del 1979 e Horace R. Drew nel nostro laboratorio determinò la struttura del tetramero *CGCG* pochi mesi dopo. Con sorpresa di tutti, le due molecole non risultarono né DNA *A* né DNA *B*. Non erano neppure delle eliche con avvolgimento (o avvitamento) destrorso, ma al contrario erano eliche con avvolgimento sinistrorso di un tipo interamente nuovo e con una particolare impalcatura a zig zag, che fece sì che venissero denominate eliche *Z*. Gli studi sulle fibre ci avevano preparato a porci un interrogativo ben definito: queste molecole si trovano nella forma *A* o nella forma *B*? La natura ci dava una risposta quasi rude e senza scampo: niente di tutto ciò.

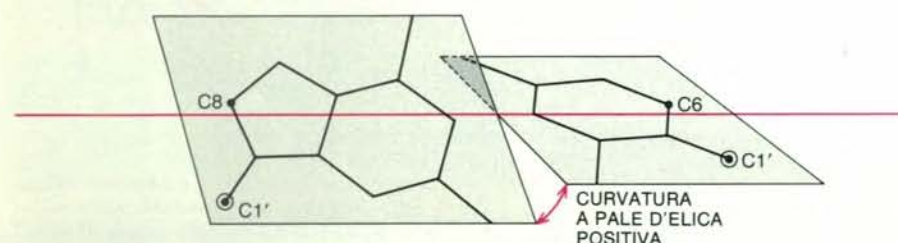
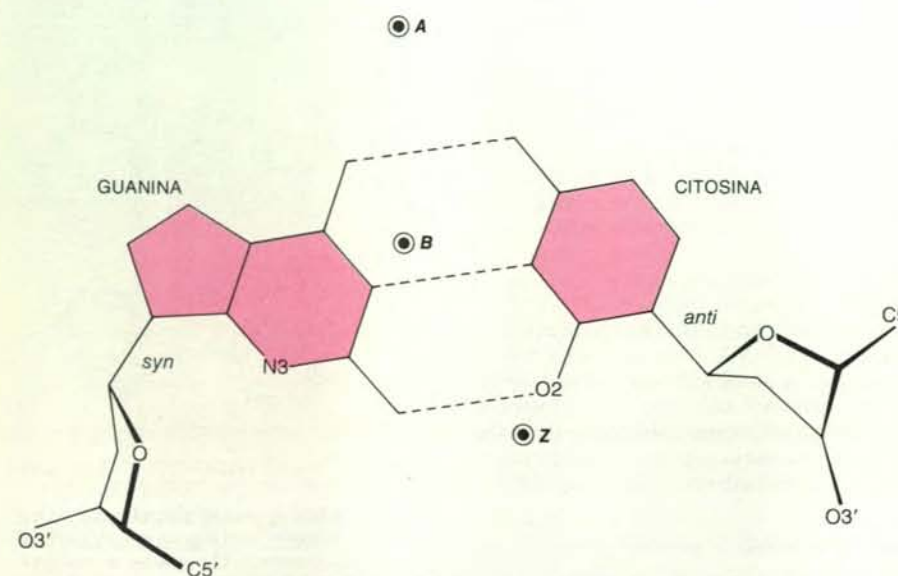
Benché le prime due molecole di DNA a doppia elica, risolte con i metodi di diffrazione dei raggi X in cristalli singoli, avessero deluso tutte le attese, le successive due molecole fornirono ciascuna un esempio di DNA *B* e *A*. Drew sintetizzò il dodecamero *CGCGAATTCGCG*, un polimero che scegliemmo per due ragioni: esso univa estremità *CGCG Z*-compatibili con un centro *AATT Z*-incompatibile, fornendo così una prova del potere di formazione dell'elica *Z* da parte di *CGCG* in un ambiente estraneo. Il dodecamero includeva inoltre la sequenza *GAATTC*, che era interessante in quanto sito di riconoscimento e di scissione a opera dell'*Eco RI*, una delle endonucleasi di restrizione, enzimi che tagliano le molecole di DNA in corrispondenza di siti specifici. (Il lettore potrà notare che l'oligomero *CGCGAATTCGCG*, come altri sintetizzati per questi studi, è complementare di se stesso, nel senso che la sequenza letta da sinistra a destra è complementare di quella letta da destra a sinistra, per cui due qualsiasi di queste catene possono combinarsi e formare una doppia elica. Ne risulta un importante risparmio di tempo e di energia, poiché deve essere sintetizzato solo uno dei filamenti della doppia elica.) La struttura della molecola di *CGCGAATTCGCG*, apparve totalmente estranea al DNA *Z*, ma fornì piuttosto un esempio classico di elica *B*. Benjamin N. Conner nel nostro laboratorio ha anch'egli sintetizzato il tetramero *CCGG*, ne ha risolto la struttura e ha mostrato che si tratta di un segmento di DNA *A*.

Altre molecole di DNA a doppia elica hanno fatto seguito a queste, in rapida successione, e sono state prodotte da parecchi laboratori. L'elica *A* viene rappresentata dal *CCGG*, dal *GGCCGGCC* del gruppo del MIT e dal *GGTATACC* di un gruppo che include Olga Kennard dell'Università di Cambridge e Zippora Shakked del Weizmann Institute of Science in Israele. Wang ha anche risolto la struttura di una molecola di DNA *A*, che è una doppia elica ibrida mista di RNA e di DNA: (*GCG*)*TATACGC*. (Le lettere racchiuse tra parentesi rappresentano le basi dell'RNA, dimodoché le tre coppie di basi più esterne, a ogni estremità della doppia elica, sono ibridi RNA-DNA e le quattro coppie di basi centrali sono DNA puro.) Il DNA *B* è rappresentato finora solo dal *CGCGAATTCGCG* e dai suoi derivati. Il DNA *Z* si trova nei già menzionati *CGCG* e *CGCGCG* e in una variante di quest'ultimo in cui gruppi metilici vengono aggiunti alle molecole di citosina: una modificazione, questa, che favorisce la struttura *Z*. Si possono oggi analizzare queste molecole per vedere quanto esse concordino con le previsioni fatte per il DNA *A* e per il DNA *B* e ricavate da studi sulla diffrazione nelle fibre, quanta variazione esista rispetto alle proprietà medie dell'elica e in quale misura la variazione possa essere attribuita alla sequenza di basi dello stesso DNA.

Le caratteristiche strutturali

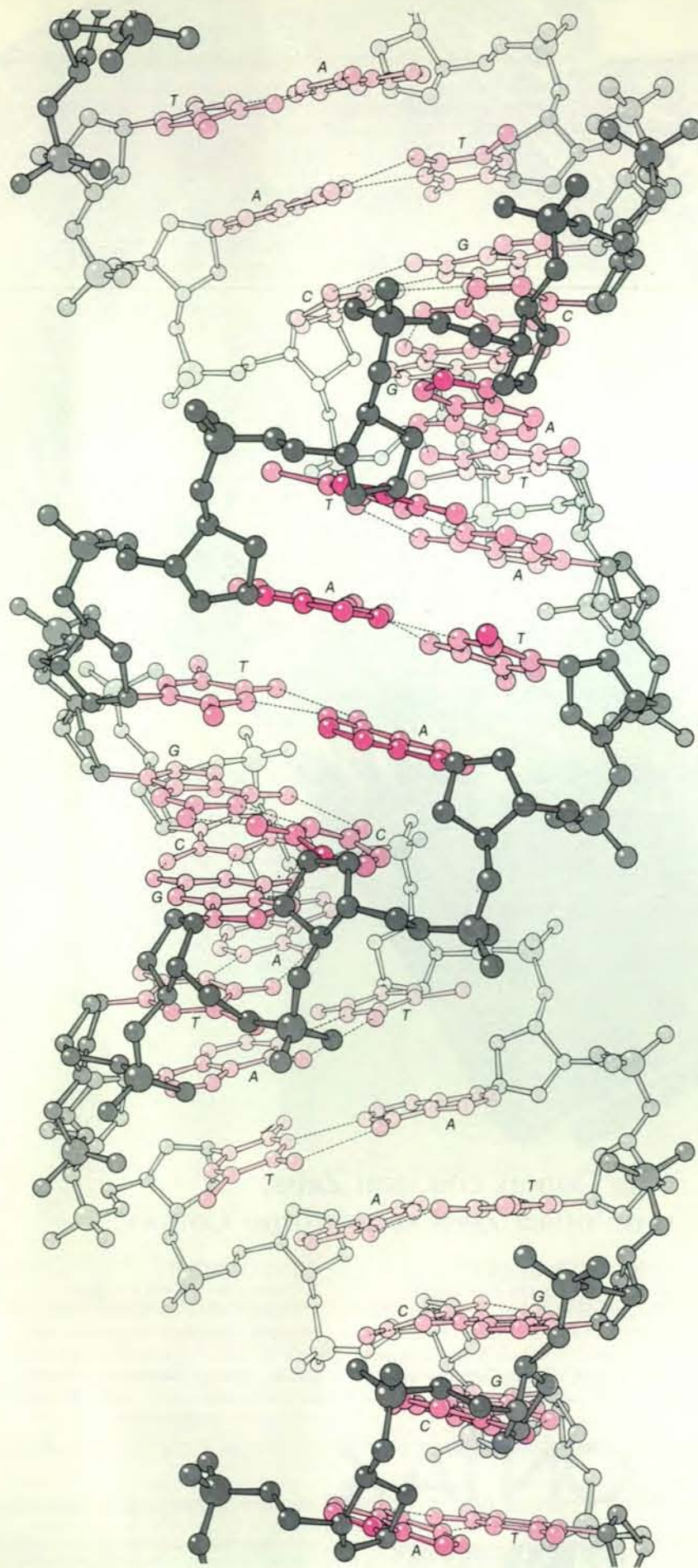
I valori medi per le caratteristiche strutturali di ogni tipo di elica sono stati determinati in base ad analisi compiute su cristalli singoli (*si veda la tabella nella pagina successiva*). Le molecole di DNA *A* e *B* hanno un avvitamento destrorso, con angoli di avvolgimento elicoidale medi, da una coppia di basi alla successiva, rispet-

L'elica è qui srotolata per evidenziare in maniera schematica i due montanti zucchero-fosfato e i pioli costituiti dalle coppie di basi. I due montanti decorrono in direzioni opposte, con le estremità 5′ e 3′ così chiamate per l'orientamento degli atomi di carbonio in posizione 5′ e 3′ negli anelli dello zucchero. Ogni coppia di basi ha una purina, adenina (*A*) o guanina (*G*), e una pirimidina, timina (*T*) o citosina (*C*), unite da legami a idrogeno (*punteggiato*).

Le coppie di basi appaiono qui nei particolari e i tratteggi indicano i loro legami con gli anelli di zucchero dei montanti. Ognuna ha un margine (verso l'alto) che delimita il solco maggiore e uno (verso il basso) che delimita il solco minore. I gruppi NH_2 sono potenziali donatori di idrogeno nella formazione di legami a idrogeno con i repressori o altre molecole regolatrici; l'azoto e l'ossigeno, in colore, ne sono potenziali accettori. In una coppia di basi, una distribuzione tipica di donatori e accettori viene probabilmente «letta» da proteine di controllo.

	DNA *A*	DNA *B*	DNA *Z*
AVVITAMENTO	DESTRORSO	DESTRORSO	SINISTRORSO
AVVOLGIMENTO ELICOIDALE (GRADI)			
MEDIA E DEVIAZIONE STANDARD	33,1 ± 5,9	35,9 ± 4,3	*G–C:* −51,3 ± 1,6 *C–G:* −8,5 ± 1,1
AMBITO DI OSSERVAZIONE	da 16,1 a 44,1	da 27,7 a 42,0	
COPPIE DI BASI PER GIRO	10,9	10,0	12,0
INNALZAMENTO DELL'ELICA PER COPPIA DI BASI (NANOMETRI)	0,292 ± 0,039	0,336 ± 0,042	*G–C:* 0,352 ± 0,022 *C–G:* 0,413 ± 0,018
INCLINAZIONE DELLE BASI (GRADI)	13,0 ± 1,9	−2,0 ± 4,6	8,8 ± 0,7
CURVATURA A PALE D'ELICA DELLE COPPIE DI BASI (GRADI)	15,4 ± 6,2	11,7 ± 4,8	4,4 ± 2,8
OSCILLAZIONE DEL PIANO DELLE BASI (GRADI)	5,9 ± 4,7	−1,0 ± 5,5	3,4 ± 2,1

Sono qui elencati, per tre tipi di DNA, i parametri medi dell'elica e le deviazioni standard. Questi dati si basano sull'analisi ai raggi X di cristalli singoli di molecole menzionate in questo articolo.

La curvatura a pale d'elica è definita da questo disegno schematico, riferito a una coppia di basi purina-pirimidina. Una rotazione in senso orario della base più vicina allorché si guarda da una direzione o dall'altra verso l'asse maggiore che attraversa la coppia di basi, viene considerata positiva. Si noti che gli atomi C1′, mediante i quali le basi sono attaccate alle impalcature di sostegno zucchero-fosfato, sono spostati in sù o in giù da questa piegatura a pale d'elica.

Vengono qui illustrate due conformazioni del legame C–N, che connette ogni base al rispettivo anello di zucchero. La conformazione *anti* (*a destra*) compare in tutti i DNA *A* e *B* e in corrispondenza delle citosine nel DNA Z. La conformazione *syn* (*a sinistra*) compare, invece, in corrispondenza delle guanine nel DNA Z. L'alternanza di una purina e di una pirimidina (*G, C*) lungo un filamento di DNA *Z* rende possibile l'alternanza *syn-anti*, che determina l'andamento a zig-zag dell'elica *Z*. I tre punti contrassegnati con *A, B* e *Z* indicano, per ogni tipo di elica, la posizione dell'asse nei riguardi di una coppia di basi. L'asse si trova nel solco maggiore nel DNA *A*, passa attraverso la coppia di basi nel DNA *B* e si trova, infine, nel solco minore nel DNA Z.

tivamente di 33,1 e 35,9 gradi, il che corrisponde a 10,9 e 10 coppie di basi per un giro completo di 360 gradi, in stretta concordanza con le previsioni fatte sulla base degli studi sulle fibre. Anche l'innalzamento (o la progressione) lungo l'asse dell'elica per ogni coppia di basi non presenta sorprese. Ciò che sorprende è l'entità della variazione nei valori degli angoli di avvolgimento a elica rispetto a questi valori medi: una deviazione standard di ± 6 gradi per il DNA *A* e di ± 4 gradi per il DNA *B*. I singoli angoli di avvolgimento sono o di soli 16 gradi o perfino di 44 gradi nel DNA *A*, mentre variano tra 28 e 42 gradi nel DNA *B*. Questa variazione nell'avvolgimento elicoidale locale può essere prevista direttamente partendo dalla sequenza delle basi con un metodo che descriverò più avanti. (La suddetta variabile è forse un elemento mediante il quale il repressore e altre molecole di controllo riconoscono particolari sequenze di basi? L'idea è attraente.) Le basi sono inclinate, rispetto al piano perpendicolare all'asse dell'elica, grosso modo di quel valore che era stato previsto in base all'analisi delle fibre: 13 gradi nel DNA *A* e circa −2 gradi nel DNA *B*.

Un maggior allontanamento dai valori attesi in base alle analisi sulle fibre si trova nel caso della cosiddetta «curvatura a pale d'elica» di singole coppie di basi. Tale curvatura o piegatura è, in pratica, una rotazione delle due basi di una coppia in direzioni opposte attorno al loro asse lungo: il senso della piegatura è definito positivo quando la base più vicina (guardando lungo quest'asse) è ruotata nel senso orario (*si veda l'illustrazione al centro qui a sinistra*). Una scarsa attenzione era stata rivolta, nello studio delle fibre, alla curvatura a pale d'elica. Una serie di coordinate per il DNA *A*, pubblicata nel 1972, contemplava una curvatura negativa, mentre una revisione del 1981, però non pubblicata, contemplava una curvatura più consona, ma con una dimensione dell'angolo di soli otto gradi.

Gli studi sui cristalli singoli mostrano che la curvatura a pale d'elica è sempre positiva, con valori medi di 15 gradi per il DNA *A* e di 12 gradi per il DNA *B*. I singoli valori si collocano entro una gamma che va da soli tre gradi ad anche 25 gradi. In un'elica destrogira, la ripiegatura a pale d'elica positiva migliora l'impaccamento delle basi lungo ogni singolo filamento che funge da impalcatura, facendo in modo che ogni base si sovrapponga meglio alle basi sue vicine, sia da una parte sia dall'altra del filamento. D'altronde, una ripiegatura a pale d'elica positiva porta anche a uno stretto e disagevole contatto le purine (*G* e *A*) situate in corrispondenza di coppie di basi adiacenti su filamenti opposti. Questo fatto, come ha suggerito C. R. Calladine di Cambridge, è principalmente responsabile di parecchie importanti variazioni nella struttura dell'elica dipendenti dalla sequenza di basi, tra cui la stessa curvatura a pale d'elica, il locale avvolgimento, o avvitamento, a elica e quello che è chiamato l'«angolo di oscillazione» delle basi.

L'oscillazione (del piano o della coppia) delle basi dà una valutazione dell'orientamento della coppia di basi come un tutto (il miglior piano medio che passa attraverso la purina e la pirimidina) attorno al suo asse maggiore. Se due coppie di basi successive lungo l'elica sono fatte ruotare in direzioni opposte, aprono tra di esse un angolo orientato o verso il solco maggiore o verso il solco minore. L'angolo di oscillazione da una coppia di basi alla successiva viene definito positivo se si apre in direzione del solco minore, mentre esso è negativo se si apre in direzione del solco maggiore. Non vi è alcuna possibilità di misurare questi angoli locali di oscillazione delle basi partendo dai dati di diffrazione ottenuti con fibre, mentre gli studi su cristalli singoli mostrano che l'oscillazione delle basi è uno dei parametri importanti dell'elica. Dato che nel DNA *B* le coppie di basi sono pressoché perpendicolari all'asse dell'elica, non è necessario che vi sia un angolo di oscillazione medio, sistematico, da una coppia di basi alla successiva. Nella forma *A*, invece, la scala a chiocciola costituita dal DNA si avvolge attorno all'asse dell'elica in un modo che rende necessaria l'apertura delle coppie di basi verso il solco minore in forma di fisarmonica, con l'angolo medio di oscillazione di sei gradi. Ancora più significativa risulta la variazione, sia nell'elica *B* sia nell'elica *A*, di circa ± 5 gradi da questi valori medi di angolo di oscillazione.

L'espansione verso eliche infinite

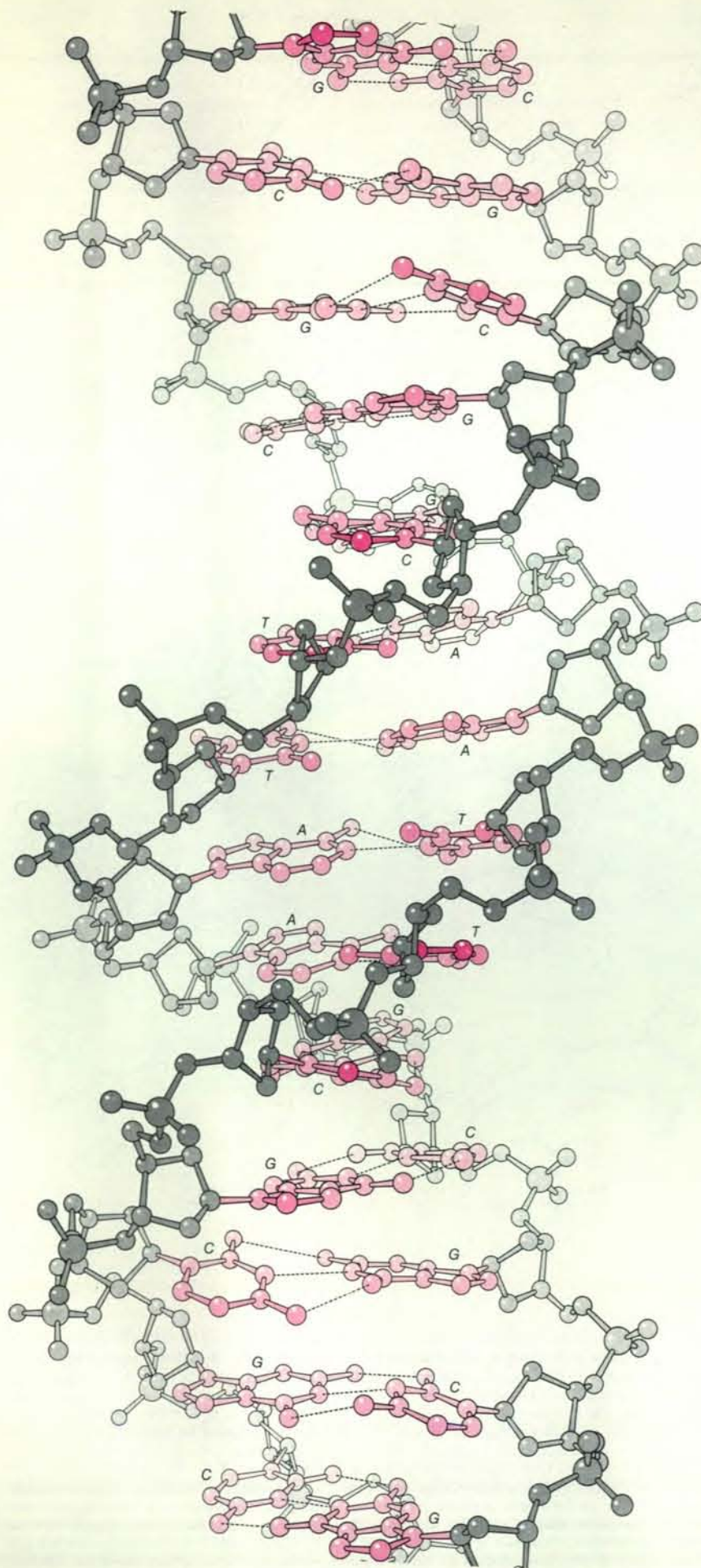
La più lunga doppia elica di DNA che sia stata finora esaminata con metodi diffrattometrici su cristalli singoli è il dodecamero *CGCGAATTCGCG*. Un'impressione dell'aspetto che avrebbero segmenti d'elica di maggiore lunghezza si può ricavare estendendo tre molecole tipiche (*si veda la figura a pagina 66*) così da produrre eliche lunghe (*si vedano le figure da questa pagina alla pagina 77*). Tutto questo si riesce a realizzare programmando un calcolatore a ruotare un'immagine di una determinata molecola a elica lungo il suo asse fino a che gli atomi vicino all'inizio dell'immagine ruotata coincidono con gli atomi equivalenti che si trovano vicino alla fine dell'immagine originale, e continuando l'operazione quanto si vuole. In questo modo, le sei coppie di basi centrali del segmento di elica *A GGTATACC* sono state ripetute quattro volte per produrre un'elica costituita da 24 coppie di basi, con la sequenza *GTATACGTATACGTATACGTATAC*. Per la prima volta è stato possibile generare una doppia elica lunga partendo da molecole ottenute mediante analisi con i raggi X su cristalli singoli, senza idealizzare fluttuazioni locali, indotte dalla sequenza di basi, nella struttura dell'elica, o farne la media.

Come previsto dagli studi sulle fibre, l'elica *A* è corta e tozza, con un solco maggiore profondo e un solco minore largo e poco profondo. L'elica *B* è più sottile (e più lunga a parità di numero di coppie di basi), con un solco maggiore largo e un solco minore stretto, di profondità paragonabile. Infine, l'elica *Z* levogira ottenuta di recente è sottile e allungata, con un solco minore profondo e stretto e un «solco» maggiore che, in realtà, è spinto alla superficie per cui non è più, di fatto, un vero solco. Dato che nel DNA *Z* il filamento a elica costituito dai residui di desossiribosio e dai gruppi fosfato segue un andamento a zig-zag, l'unità ripetitiva dell'elica non è una singola coppia di basi, come nel DNA *A* e *B*, ma piuttosto due successive coppie di basi: *G–C* seguita da *C–G*. Ciò risulta a causa di una differenza nel modo in cui le citosine e le guanine sono attaccate, nel DNA *Z*, ai rispettivi anelli di desossiribosio.

In corrispondenza di ogni citosina, il desossiribosio è ruotato attorno al legame che ha con la base, in modo che l'anello di cui è costituito e che si presenta increspato oscilli allontanandosi dal solco minore (*si veda l'illustrazione in basso a pagina 70*). Questa conformazione, designata come *anti*, è presente anche nel caso di tutte e tre le altre basi, e pertanto a ogni piolo della scala a chiocciola, che rappresenta le molecole sia del DNA *A* sia del DNA *B*. Ogni residuo di guanina nel DNA *Z*, invece, è legato con un anello di desossiribosio che è ruotato di 180 gradi, per cui si piega verso l'interno in direzione del solco minore. Questa conformazione *syn* è possibile da un punto di vista stereochimico solo quando il desossiribosio si lega all'anello più piccolo, a cinque atomi, di una purina, invece che all'anello a sei atomi di una pirimidina: la distanza tra un anello *syn* di desossiribosio e l'atomo di ossigeno in posizione O2 di un anello di citosina o di timina sarebbe troppo breve per essere accettabile da un punto di vista stereochimico. È l'alternanza tra una purina (*G*) e una pirimidina (*C*) lungo ciascun filamento del DNA *Z* che permette un'alternanza *syn-anti* dei legami tra i residui di desossiribosio e le basi e produce il filamento a zig-zag. Il percorso di questo filamento da un atomo di fosforo al successivo, dopo ogni residuo di guanina, è quasi parallelo all'asse dell'elica; d'altra parte, il percorso da un fosforo al successivo, dopo ogni residuo di citosina, è tangenziale all'involucro cilindrico dell'elica e perpendicolare al suo asse. Come risultato, la vera unità ripetitiva lungo l'elica è costituita da due coppie di basi successive: anche se l'elica *Z* ha 12 coppie di basi per giro, essa è formalmente un'elica levogira a sei componenti, vale a dire con sei gruppi di due coppie di basi per giro.

In questo disegno in prospettiva, che si basa sulla duplicazione stereo, generata al calcolatore, della pagina a fronte, si nota un'elica di DNA *A*. Tale duplicazione è stata ottenuta estendendo le sei basi centrali dell'ottamero *GGTATACC*, la cui struttura è stata determinata da Olga Kennard, Zippora Shakked e M. A. Viswamitra. Le coppie di basi non sono idealizzate, ma sono come se fossero osservate nell'analisi della struttura cristallina. Si noti come i gruppi fosfato su filamenti opposti si fronteggino da una parte all'altra del solco principale.

Questa duplicazione stereo è stata generata manipolando la serie tridimensionale di coordinate che rappresentano un'immagine dell'esamero *GTATAC* del DNA *A*, ottenuto per delezione di coppie di basi alle estremità dell'ottamero. L'immagine è stata ruotata e traslata lungo l'asse dell'elica fino a quando la coppia di fosfati in basso nell'immagine ha coinciso con la coppia in alto di un'immagine che non aveva subito spostamenti. Il processo è stato ripetuto molte volte per dare ciò che, di fatto, è un'elica *A* infinita. Il metodo permette di generare un lungo filamento di un DNA specifico senza sacrificare le variazioni locali nella struttura dell'elica, che sono determinate dalla sequenza di basi. Una duplicazione stereo si osserva nel migliore dei modi in prospettiva attraverso un visore stereo con lente d'ingrandimento. Con la pratica, un'immagine tridimensionale si riesce a vedere anche senza un visore, se si riesce a sviluppare la capacità di disaccoppiare due riflessi che sono normalmente associati: divergere gli occhi come se guardassero un oggetto lontano mettendo contemporaneamente a fuoco gli occhi per la visione vicina. Si noti la profondità del solco maggiore. L'elica assomiglia a un nastro duplice, avvolto ad elica all'esterno di un cilindro.

L'elica del DNA *B* illustrata qui è stata generata dalla ripetizione delle 10 coppie di basi centrali del dodecamero *CGCGAATTCGCG*. È evidente la marcata piegatura a pale d'elica del piano delle coppie di basi, come pure il modo in cui tale curvatura favorisce la sovrapposizione delle basi in ogni filamento dell'elica.

L'RNA

Finora ho parlato soltanto del DNA (acido desossiribonucleico), il materiale presente nel nucleo della cellula e che funge da archivio per l'informazione genetica. L'espressione di quest'ultima dipende in grande parte da un altro acido nucleico, molto affine al primo, l'acido ribonucleico o RNA. La sequenza di basi su un filamento dell'elica di DNA viene trascritta in un unico filamento di RNA messaggero: le molecole di RNA di trasporto leggono quindi quest'ultima sequenza e portano ai ribosomi, dove essa si trova, gli amminoacidi designati perché vengano riuniti in una catena polipeptidica. Gli stessi ribosomi sono costituiti in parte da RNA. Vi sono appena due differenze tra DNA e RNA. La timina del DNA viene sostituita nell'RNA dall'uracile che manca della catena laterale metilica della prima base; inoltre, al posto degli anelli di desossiribosio del DNA, l'impalcatura di sostegno dell'RNA ha anelli di ribosio. Nel desossiribosio, il gruppo ossidrilico (OH), legato a un atomo di carbonio (C2′) dell'anello di ribosio, viene sostituito da un idrogeno.

Un'importante conseguenza della presenza di un ossidrile supplementare nell'acido ribonucleico è che l'RNA chiaramente non riesce ad avvolgersi in una doppia elica *B*. In un ipotetico RNA *B* il gruppo OH sarebbe localizzato al centro di una gabbia di atomi appartenenti al gruppo fosfato, all'anello di ribosio e alla base. La distanza tra l'atomo di ossigeno supplementare e i diversi altri atomi sarebbe troppo breve e renderebbe pertanto la struttura piuttosto proibitiva da un punto di vista stereochimico. D'altra parte, nel DNA *A* il gruppo OH in posizione 2′ sporge dalla superficie dell'elica verso l'esterno, allontanandosi da ogni atomo vicino.

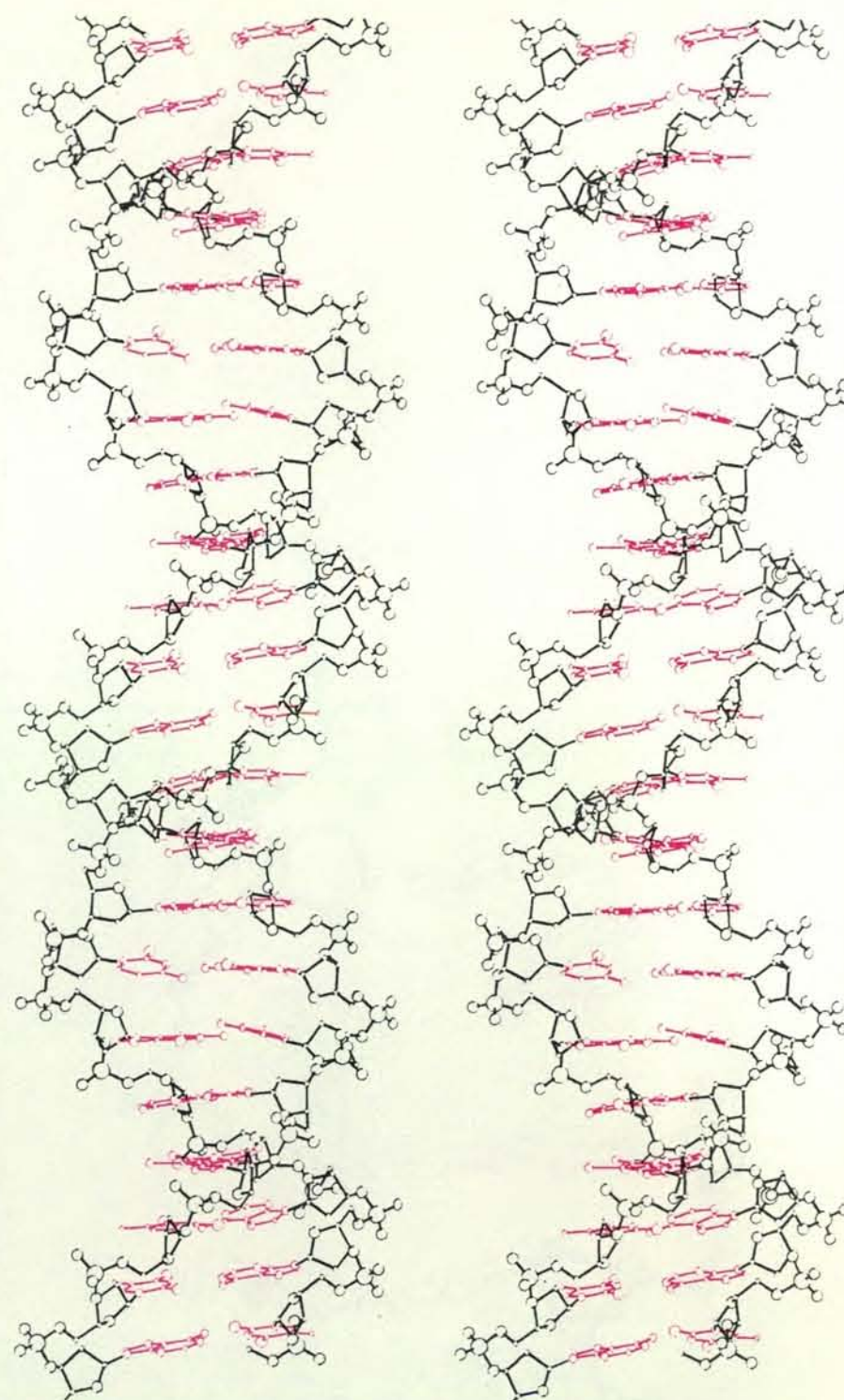
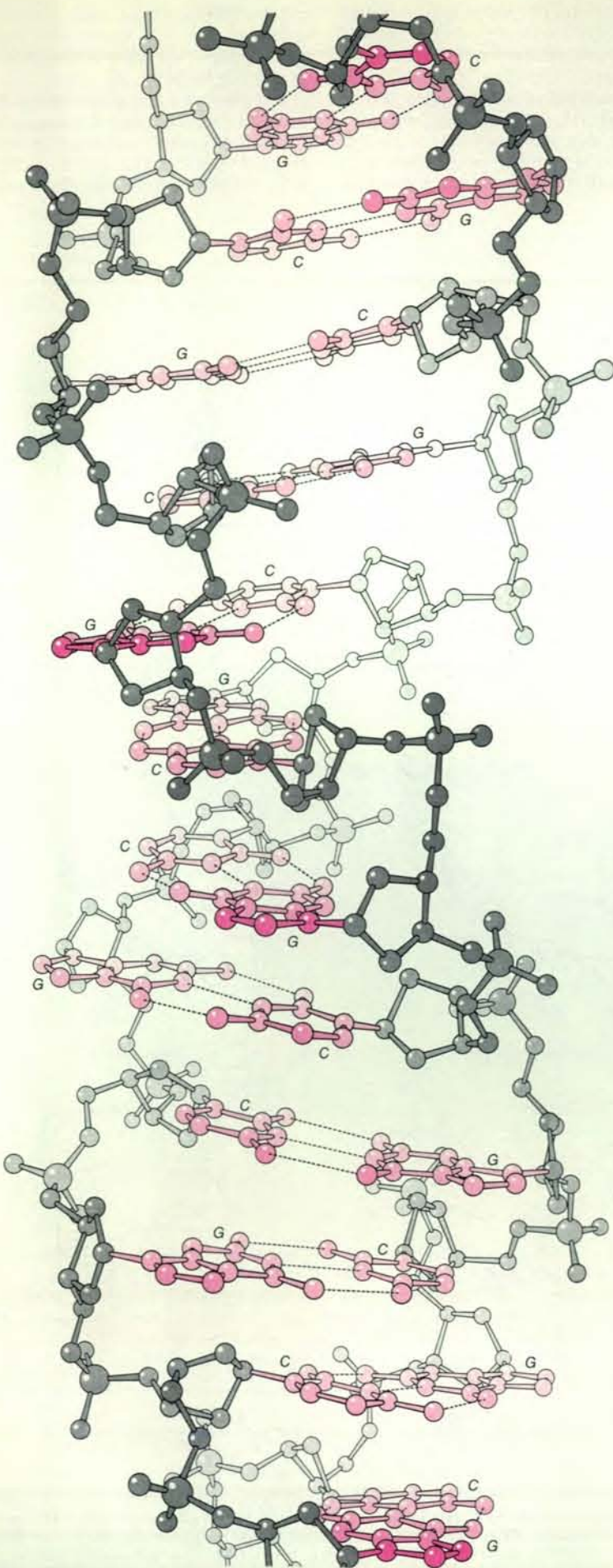
Nella molecola di RNA di trasporto le anse e gli avvolgimenti a doppia elica, complementari di sé, devono essere pertanto variazioni di un'elica *A*. Se, come è stato suggerito, brevi tratti complementari di sé dell'RNA messaggero a filamento singolo possono formare, mediante legami a idrogeno, ripiegature a forcina, anche queste ultime devono essere RNA *A* invece di *B*. (Nell'unico esempio noto di partecipazione dell'RNA in un'elica *B*, un ibrido sintetico con residui di adenina sul filamento di RNA e di timina su quello di DNA, l'OH supplementare si collega mediante un legame a idrogeno all'atomo di ossigeno in posizione 4′ del successivo ribosio.) Nessuno ha seriamente suggerito la possibilità di esistenza di un RNA *Z*, ma un attento esame della struttura indica che un OH aggiunto a un nucleoside citidilico probabilmente verrebbe a trovarsi troppo vicino all'atomo di ossigeno in posizione 2 dell'anello della citosina, che sporge.

Stabilizzazione per idratazione

L'analisi strutturale mediante l'impiego di raggi X su cristalli singoli ha definito tre tipi fondamentali di doppia elica di DNA, invece di due, e un tipo di doppia elica di RNA. Per il DNA, la forma *B* è stabile in condizioni di elevata umidità e questo fa sembrare che vi sia una probabilità massima per l'elica *B* di essere presente nel nucleo cellulare. Se un'elica ibrida transitoria DNA-RNA si forma nel corso della trascrizione del DNA in RNA messaggero, è probabile che essa adotti la struttura *A*. In speciali condizioni di tensione, il DNA destrogiro, con la sequenza in cui si alternano purine e pirimidine, può essere tramutato nello stato *Z* levogiro. Invece, in sequenze di carattere generale e in condizioni di elevata umidità, il DNA *B* sembra costituire la norma. In che modo l'acqua rende stabile l'elica *B*?

Dopo che la struttura di una molecola di DNA è stata precisata mediante analisi ai raggi X, si può procedere a una prudente ricerca sulle molecole di solvente (acqua) che la circondano e che risultano estremamente ordinate, cioè così fortemente legate al DNA da occupare in ogni molecola del cristallo sempre gli stessi posti. (Molecole d'acqua meno strettamente legate, che si trovano in punti diversi in una molecola rispetto alla molecola adiacente, all'interno del cristallo, sono indistinte nei diffrattogrammi e sono scarsamente individuabili rispetto al rumore di fondo.) La struttura dell'acqua è stata esaminata nelle molecole di tutte e tre le forme della doppia elica di DNA, ma risultati particolareggiati sono stati pubblicati solo per il DNA *B*. In quest'ultimo, le molecole d'acqua si osservano in vicinanza di quasi ogni atomo che potrebbe formare con loro dei legami a idrogeno: ossigeni liberi dei gruppi fosfato, atomi di azoto e di ossigeno sui margini delle coppie di basi e, in minor misura, gli ossigeni dei gruppi fosfato che fanno parte dell'impalcatura di sostegno. Di fatto l'elica è rivestita da uno strato d'acqua dello spessore di una molecola.

Una struttura con una idratazione più estesa si osserva all'interno dello stretto solco minore del DNA *B* (*si veda la figura in basso a pagina 78*). La combinazione della curvatura a pale d'elica di singole coppie di basi e dell'avvolgimento, o avvitamento, elicoidale da una coppia di basi alla successiva porta un atomo di ossigeno o di azoto di un filamento in stretta prossimità di un atomo di ossigeno o di azoto dell'altro filamento sulla coppia di basi adiacente. I due accettori del legame a idrogeno sono così uniti da una molecola d'acqua. Queste molecole del primo strato sono unite a loro volta da un secondo strato di molecole d'acqua. Assieme i due gruppi di molecole costituiscono una specie di «spina dorsale» idratata che decorre lungo il solco minore, avvolgendosi attorno all'elica *B*. Questa spina dorsale è ben sviluppata nelle regioni con coppie di basi *A–T*, mentre è interrotta nelle regioni con coppie di basi *G–C*, probabilmente a causa dell'intrusione nel solco minore dei gruppi NH_2 delle guanine. Nelle fibre di DNA sintetico qualunque coppia di basi in cui la guanina sia sostituita dalla base modificata inosina, che è priva del gruppo NH_2 della guanina, si comporta come una coppia *A–T* piuttosto che come una coppia *C–G*, nel senso che favorisce la stabilità della struttura *B*.

Noi riteniamo che questa specie di spina dorsale idratata del solco minore abbia un'influenza stabilizzatrice di prim'ordine sul DNA *B* e che debba fisicamente rompersi prima che il solco minore si pos-

Questa duplicazione stereo è stata preparata eliminando le basi più esterne del dodecamero e reiterando la sequenza *GCGAATTCGC*. In un'elica *B* i solchi maggiore e minore hanno all'incirca la stessa profondità, anche se il minore è in realtà più stretto; lo è particolarmente nella regione *AATT* in avanti, mentre è più ampio posteriormente, nelle regioni ricche di coppie di basi *C–G*.

sa aprire e che l'elica passi dalla forma *B* alla forma *A*. Un modo di distruggere in laboratorio la spina stabilizzatrice consiste nel rimuovere le molecole d'acqua disidratando le fibre. Un altro modo consiste invece nell'introdurre più gruppi NH_2 nel solco minore: studi diffrattometrici sulle fibre, effettuati da Struther Arnott della Purdue University, hanno mostrato che l'aumento del contenuto di *G–C* in una fibra facilita la transizione dalla forma *B* alla forma A per essiccamento.

Nessuna struttura sistematica per l'acqua, paragonabile a questa, è stata trovata attorno all'elica *A* o all'elica *Z*. Anche se le molecole d'acqua sono distribuite liberamente attorno ad atomi sul DNA che potrebbero prendere parte ai legami a idrogeno, non vi è nulla che sia confrontabile con la spina dorsale di molecole d'acqua, presente nel solco minore del DNA *B*, e che potrebbe presumibilmente avere una integrità strutturale. Di fatto, il largo solco minore di un DNA *A* sembra essere addirittura meno idratato di altre parti dell'elica. Tuttavia, è difficile oggi essere sicuri dell'idratazione del DNA *A*, dato che in tutti e tre i suoi cristalli - *CCGG/CCGG* (due tetrameri riuniti a formare quello che, in effetti, è un ottamero), *GGCCGGCC* e *GGTATACC* - le coppie di basi terminali di un ottamero risultano stipate contro i solchi minori di molecole vicine, bloccando in parte la normale idratazione che ci si aspetterebbe di trovare se le molecole fossero libere in soluzione. Ma anche così si può avere oggi un'idea a livello molecolare dei motivi per i quali la forma *B* prevale nelle condizioni più generali di massima umidità e si ha una transizione alla forma *A* come risultato di una disidratazione.

Il controllo genetico

Passiamo ora a quelle che possono essere le implicazioni biologiche più significative degli studi su singoli cristalli del DNA, effettuati mediante raggi X: l'influenza della sequenza di basi sulla struttura a elica locale e il ruolo che questa può svolgere nel controllare l'espressione dell'informazione genetica. Se l'elica *Z* è davvero limitata a solo un piccolo numero di sequenze di basi, con le purine e le pirimidine che si alternano, è improbabile che essa abbia molto significato nelle regioni del gene che codificano per le proteine. Tuttavia, come hanno suggerito Rich e altri, la capacità di particolari regioni regolatrici del DNA di invertire il senso del loro avvitamento elicoidale potrebbe avere un'importanza critica nel controllare la possibilità di accesso e di lettura dell'informazione genetica, per esempio modificando il grado di superavvolgimento del DNA circolare.

Un simile meccanismo costituirebbe un esempio dell'influenza che la sequenza di basi ha non sul contenuto del messaggio genetico, ma piuttosto sulla sua espressione. L'espressione selettiva o la repressione dell'informazione genetica in cellule particolari e in momenti particolari, comunque possa venir realizzata, è la chiave per spiegare lo sviluppo in forme di vita superiori, pluricellulari. Inoltre, un guasto nella regolazione dell'informazione genetica può scatenare il cancro e può ugualmente essere implicato nel processo di senescenza. L'attuale livello di comprensione del controllo dell'informazione genetica nelle forme superiori di vita è all'incirca uguale a quello che si aveva nel 1953 nei riguardi della codificazione di tale informazione, allorché Watson e Crick proposero per primi la doppia elica.

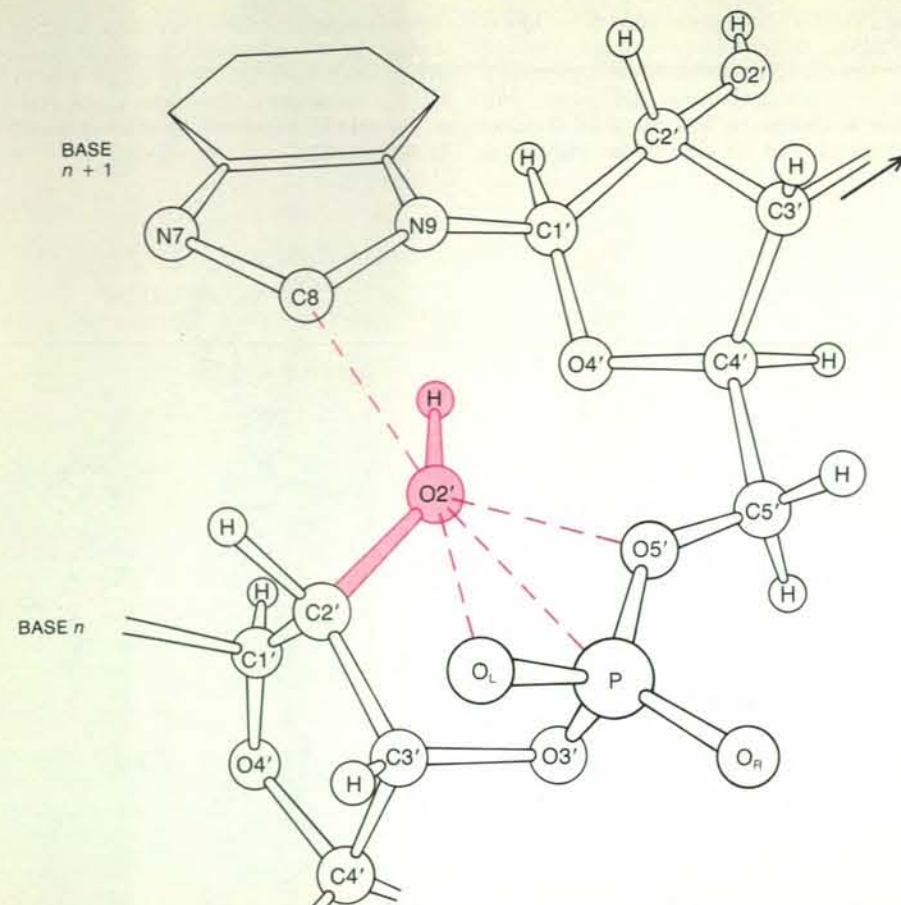
La transizione da destrorso a sinistrorso nel passaggio da DNA *B* a DNA *Z* rappresenta un grosso cambiamento conformazionale dipendente dalla sequenza delle basi. Un differente tipo di influenza strutturale sulla decifrazione dell'informazione genetica viene oggi rivelato da uno studio particolareggiato sulla variazione locale dell'elica nel DNA *A* e nel DNA *B*. Come ho indicato, i valori medi dei parametri dell'elica sono in generale pressappoco quelli che ci si attendeva di trovare dagli studi sulle fibre; ciò che non ci si attendeva di trovare, invece, erano le ampie deviazioni locali da queste medie.

L'angolo medio dell'avvolgimento elicoidale per il DNA *A* è 33,1 gradi, il che corrisponde quasi esattamente a 11 coppie di basi per giro, ma chi avrebbe previsto che i singoli valori potevano variare da 16 gradi fino a 44? Gli angoli medi di oscillazione del piano delle basi, pari a 6 gradi per il DNA *A* e a quasi zero per il DNA *B*, potevano anche essere previsti partendo dal modo in cui le coppie di basi sono avvolte attorno all'asse dell'elica, ma nessuno si aspettava di trovare delle deviazioni standard da queste medie di circa ± 5 gradi in ogni caso. Queste fluttuazioni locali fecero a tutta prima riesaminare attentamente ai cristallografi le loro analisi, così da assicurarsi che le variazioni osservate non erano semplicemente artefatti del processo di purificazione. Soddisfatto questo punto (e avendo osservato di fatto analoghe variazioni in differenti strutture purificate in tre modi diversi), la domanda ovvia fu quella di sapere se le fluttuazioni hanno qualche base prevedibile oppure no.

Dalla sequenza alla struttura

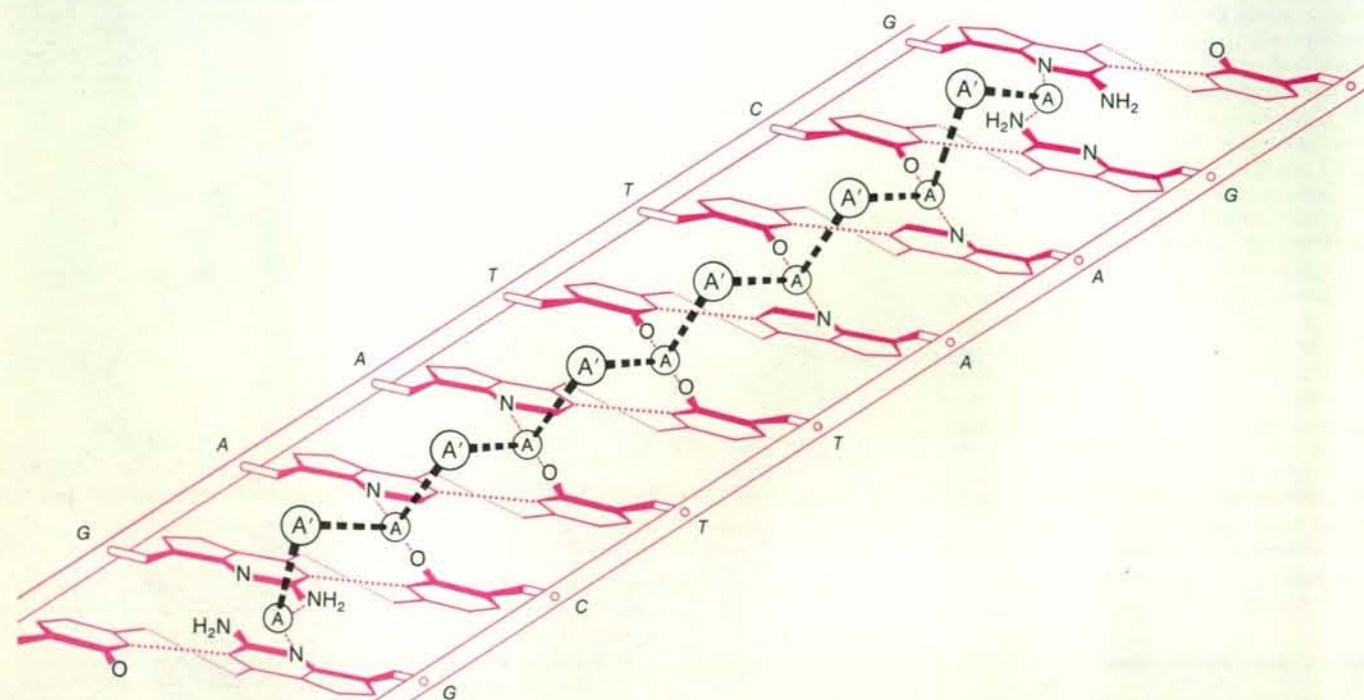
Sono stati compiuti parecchi tentativi per trovare una spiegazione alla variazione della struttura in termini di sequenza delle basi lungo il DNA. Finora il più riuscito è quello messo in opera da Calladine, il quale ha applicato alla doppia elica i principi della meccanica di un fascio elastico. Secondo l'analisi di questo autore, le variazioni locali nell'avvolgimento a elica e nell'oscillazione del piano delle basi si instaurano a causa di un impedimento sterico tra grosse basi puriniche, una conseguenza della curvatura a pale d'elica che si verifica nelle singole coppie di basi. (Si ricordi che tale piegatura favorisce il sovrapporsi delle basi lungo ciascuno dei due filamenti che costituiscono la doppia elica.)

In questo disegno, l'elica del DNA *Z* appare come una molecola ad avvitamento sinistrorso, costituita da guanine che si alternano a citosine. Essa è stata prodotta partendo dalle quattro coppie di basi centrali dell'esamero *CGCGCG*, studiato da Wang, Rich e collaboratori. La curvatura a pale d'elica è più modesta rispetto a quella che si osserva nelle due eliche ad avvitamento destrorso. I gruppi fosfato su differenti catene si trovano uno di fronte all'altro all'interno del solco minore profondo.

Questa duplicazione stereo dell'elica del DNA *Z* distesa è una reiterazione del tetramero *GCGC*. Si noti che la struttura del DNA *Z* e del DNA *A* sono, sotto molteplici aspetti, l'una l'inversa dell'altra. L'elica *Z* è lunga e sottile, con un solco minore profondo, un solco maggiore appiattito e una lieve piegatura a pale d'elica. L'elica *A* è, invece, una molecola corta e tozza, con un profondo solco maggiore, un solco minore superficiale e una notevole piegatura a pale d'elica.

L'RNA non adotta la struttura ad elica *B* quando, complementare di se stesso, forma una doppia elica. La ragione sembra essere la seguente: l'RNA presenta anelli di ribosio, con un ossidrile (OH) dove l'anello di desossiribosio del DNA ha soltanto un idrogeno. Qui un OH è stato aggiunto all'impalcatura del DNA *B* per simulare un RNA *B*. Ma tra l'atomo di ossigeno aggiunto e i quattro atomi degli adiacenti fosfato, zucchero e base si stabilisce uno scomodo contatto (*tratteggio*).

Dato che una purina a doppio anello (*G* o *A*) è più grossa di una pirimidina a singolo anello (*C* o *T*), in una qualsiasi coppia di basi essa si estende sempre oltre l'asse dell'elica. Se due purine si trovano sui due filamenti opposti dell'impalcatura del DNA, in corrispondenza di coppie di basi adiacenti, esse si sovrappongono leggermente quando sono viste in proiezione lungo l'asse della doppia elica. Se, a questo punto, a ogni coppia di basi viene conferita una curvatura a pale d'elica positiva, i margini delle purine si vengono a trovare tra loro a stretto contatto in un modo inaccettabile e ne risulta una collisione stereochimica.

Per una sequenza purina-pirimidina in progressione lungo una delle catene nucleotiche del DNA nella normale direzione 5′ → 3′, i contatti troppo stretti si hanno nel solco principale; nel caso contrario, cioè per l'alternanza pirimidina-purina, la collisione ha luogo nel solco minore (*si veda la figura nella pagina a fronte*). Come possono essere evitati questi urti tra i margini delle basi puriniche?

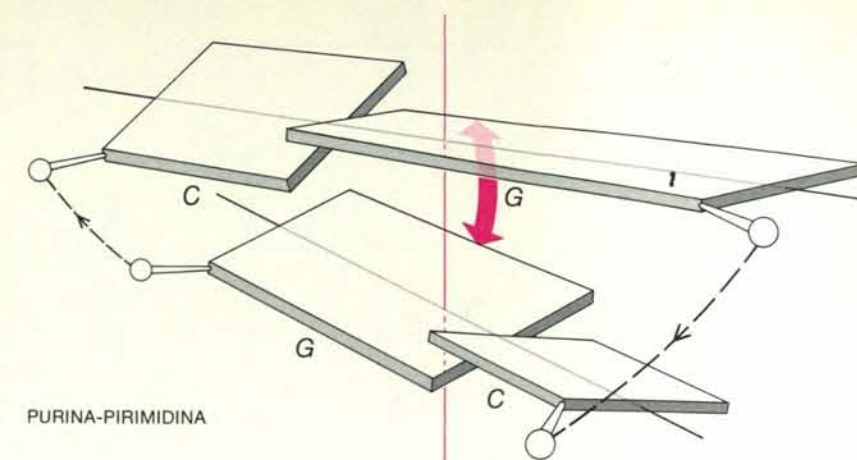
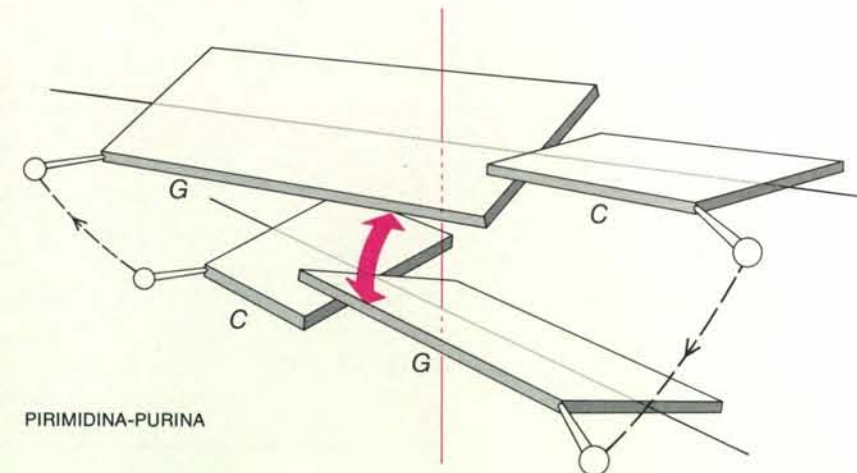
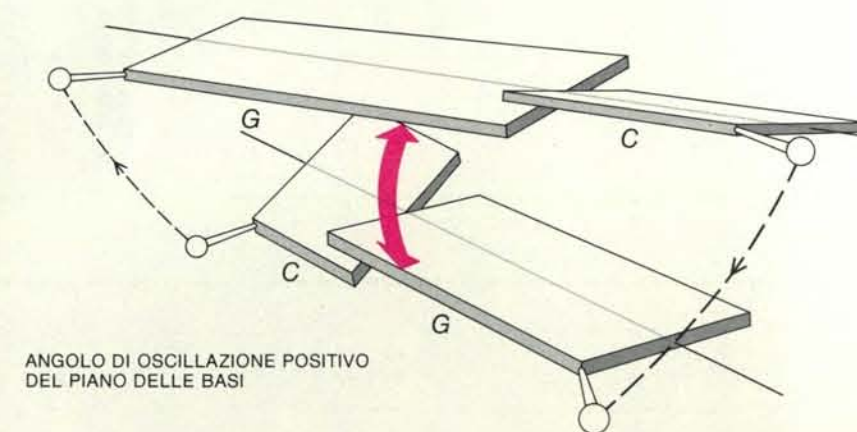
Uno dei modi consiste semplicemente nell'allargare l'angolo di oscillazione delle basi dal lato in cui i margini delle purine vengono a contatto. Nel caso della collisione che avviene nel solco minore per la alternanza pirimidina-purina, la giusta risposta sarebbe quella di allargare l'angolo tra il piano medio delle coppie di basi verso il solco minore, rendendo più positivo il valore dell'angolo di oscillazione. In corrispondenza di alternanze purina-pirimidina, tale angolo dovrebbe invece essere allargato verso il solco maggiore, il che renderebbe il suo valore più negativo. Drew e io stesso avevamo osservato questo comportamento dell'angolo di oscillazione, che dipende dalla sequenza delle basi, nella struttura *CGCGAATTCGCG*. Ma la nostra spiegazione era più involuta e meno plausibile di quella di Calladine.

Nel DNA *B*, grazie a una schiera a zig zag di molecole d'acqua, si forma una «cresta di idratazione» o «spina dorsale idratata» Appare qui l'interno del solco minore di un'elica srotolata del dodecamero del DNA *B*. L'impalcatura è fortemente schematizzata e la piegatura a pale d'elica esagerata. Uno strato sul fondo di molecole d'acqua (A) è unito mediante legami a idrogeno con gli atomi di azoto e di ossigeno presenti su coppie di basi adiacenti. Queste molecole sono poi unite da un secondo strato di molecole d'acqua (A′) che formano la spina dorsale, la quale impedisce al solco minore di allargarsi e di formare la poco profonda depressione che si nota nel DNA *A*. Pertanto si ritiene che la spina sia una fattore importante nel trasformare il DNA *B* nella struttura ad elica stabile che si ha in condizioni di elevata idratazione.

Indipendentemente dal lato di una coppia di basi lungo il quale ha luogo la collisione con una coppia vicina, questo urto può essere sempre smorzato facendo diminuire l'angolo della curvatura a pale d'elica locale tra le due coppie di basi. Questo fa pensare che la variazione che si verifica in questa rotazione potrebbe avere la stessa semplice spiegazione stereochimica proposta per l'angolo di oscillazione. Una collisione purina-purina nello spessore dell'elica può anch'essa essere attenuata direttamente riducendo in una, o in ambedue le coppie di basi, la piegatura a pale d'elica; lo stesso risultato si può ottenere indirettamente facendo slittare una coppia di basi lungo il suo asse maggiore, in modo che la purina risulti parzialmente rimossa dall'impaccamento dell'elica. Calladine ha concluso dalla sua analisi sul secondo effetto che la collisione sterica che ha luogo nel solco minore si esplica con un'intensità doppia rispetto alla collisione che ha luogo nel solco maggiore e, pertanto, richiede una messa a punto di intensità doppia dei parametri locali dell'elica.

La previsione della variazione

A partire dalle premesse di Calladine, il nostro gruppo (che oggi lavora all'Università della California a Los Angeles) ha sviluppato quattro semplici funzioni sommatorie che sono servite a prevedere, da una data sequenza di basi del DNA, le variazioni locali nell'avvolgimento a elica, nell'angolo di oscillazione delle basi, nella curvatura a pale d'elica e anche nell'angolo di torsione (una misura della conformazione della catena-impalcatura di sostegno del DNA attorno al legame C4′–C3′). Le funzioni per la curvatura a pale d'elica e per l'angolo di torsione sembrano valere solo per il DNA *B*, mentre le prime due funzioni sono valide per ambedue le eliche *B* e *A*. Quale esempio del processo di previsione, la funzione sommatoria Σ_1 per l'angolo di avvitamento a elica si genera come segue.

1. A ogni intervallo tra una coppia di basi e la successiva, si assegni un numero che rifletta la tendenza relativa a quel livello verso una riduzione dell'avvitamento a elica, in modo da ridurre la collisione stereochimica. Sia questo numero –2 per le alternanze purina-pirimidina con le loro collisioni nel solco maggiore; sia esso, invece, –4 per le alternanze pirimidina-purina con i loro contatti più marcati nel solco minore; sia infine zero nel caso delle alternanze purina-purina e pirimidina-pirimidina, che non presentano impedimento sterico.

2. Si attribuiscano numeri di segno opposto, ma di grandezza pari alla metà, alle alternanze che stanno ai due fianchi. Infatti la riduzione di un intervallo tra due coppie di basi al centro, per rotazione di ciascuna coppia, allargherà le spaziature ai

La curvatura, o piegatura, a pale d'elica influenza i contatti tra purine facenti parte di filamenti opposti, in corrispondenza di coppie di basi adiacenti. Qui le purine e le pirimidine sono schematizzate come tavole, gli atomi C1′ degli anelli di zucchero sono rappresentati da piccole sfere e i montanti sono ridotti a linee tratteggiate, con frecce che puntano nella direzione 5′→3′. Per una alternanza purina-pirimidina (*in alto*), la piegatura a pale d'elica positiva porta i margini di due purine troppo vicini nel solco maggiore (*freccia in colore*). Nel caso di una alternanza pirimidina-purina (*al centro*), questo contatto troppo stretto,tanto da essere inaccettabile, si verifica nel solco minore. Una strategia per ridurre tale collisione consiste (*in basso*) nell'allargare l'angolo di oscillazione del piano delle basi (formato dall'inclinazione di coppie di basi adiacenti intese come un tutto) laddove ha luogo la collisione, in questo caso nel solco minore.

ANGOLO DI AVVITAMENTO ELICOIDALE (GRADI)

FUNZIONE SOMMATORIA Σ_1

C – G – C – G – A – A – T – T – C – G – C – G

−4 +2 +1 −2 +1 +2 −4
+1 −2 +1 +1 −2 +1
+2 −4 +2 +2 −4 +2

Σ_1 = −3 +2 −3 +2 +1 −2 +1 +2 −3 +2 −3

G – G – C – C – G – G – C – C

+1 −2 +1 +1 −2 +1
+2 −4 +2

Σ_1 = +1 −2 +3 −4 +3 −2 +1

ANGOLO DI OSCILLAZIONE DEL PIANO DELLE BASI (GRADI)

FUNZIONE SOMMATORIA Σ_2

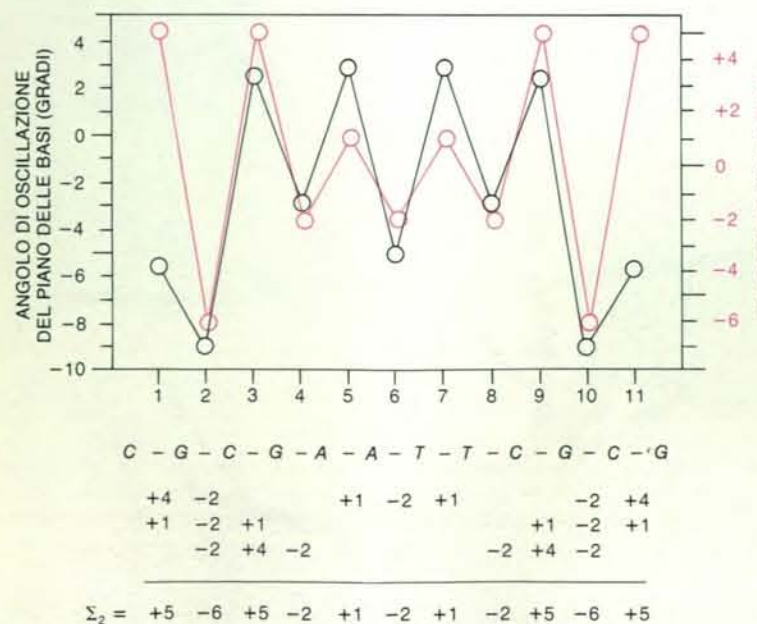
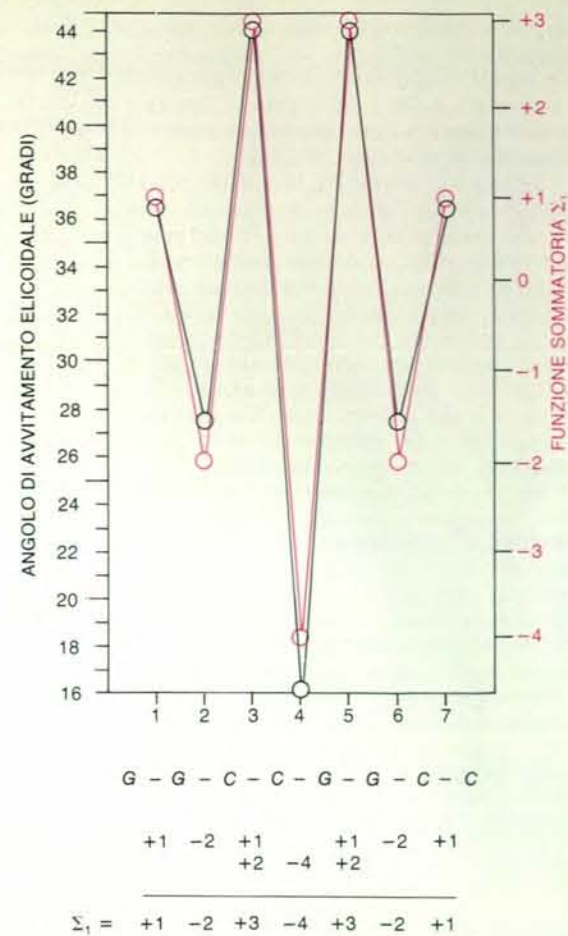
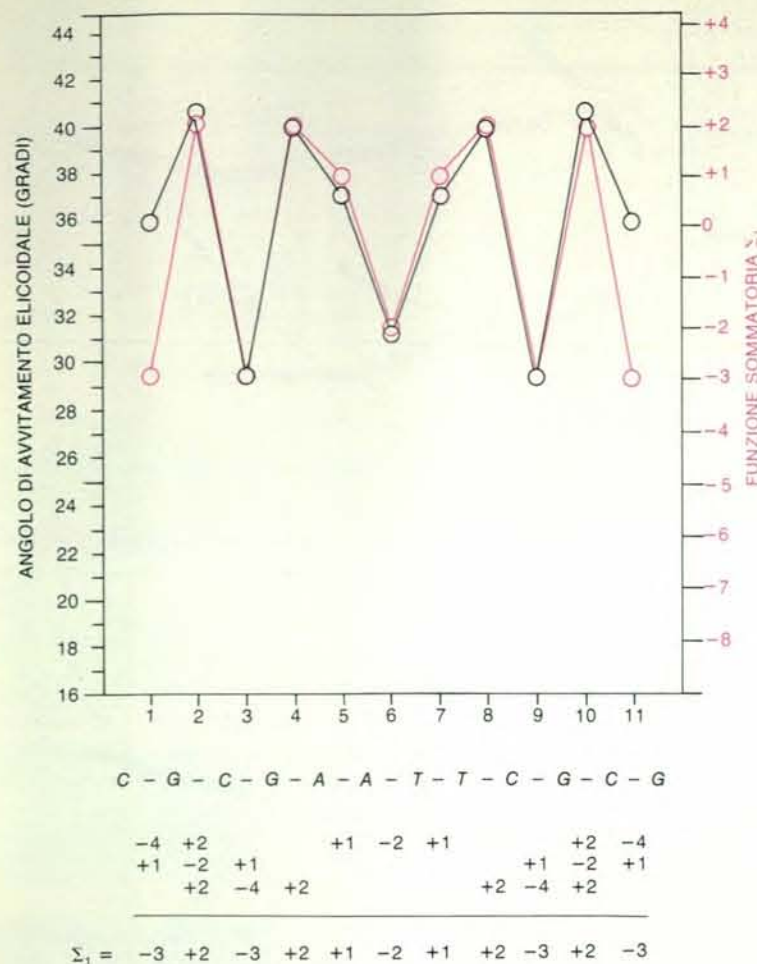
C – G – C – G – A – A – T – T – C – G – C – G

+4 −2 +1 −2 +1 −2 +4
+1 −2 +1 +1 −2 +1
−2 +4 −2 −2 +4 −2

Σ_2 = +5 −6 +5 −2 +1 −2 +1 −2 +5 −6 +5

C – C – G – G   C – C – G – G

−2 +4 −2 −2 +4 −2
+1 −2 +1

Σ_2 = −2 +4 −1 −2 −1 +4 −2

Le variazioni locali nell'angolo di avvitamento, o avvolgimento, elicoidale e nell'angolo di oscillazione del piano delle basi sono prevedibili partendo dalla sequenza di basi e ricorrendo alle funzioni sommatorie Σ_1 e Σ_2. Il metodo viene qui illustrato per il dodecamero dell'elica *B* (*in alto a sinistra e in basso a sinistra*) e per un ottamero (*in alto a destra*) e uno pseudo-ottamero (*in basso a destra*) dell'elica *A*. In ogni caso osservato, i valori sono indicati dalle curve in nero, le cui scale corrispondenti sono alla sinistra, mentre le previsioni sono indicate dalle curve in colore, le cui scale corrispondenti sono alla destra del grafico. La derivazione della funzione sommatoria per ogni alternanza tra le adiacenti coppie di basi è rappresentata sotto a ogni grafico. Le curve per i valori osservati e per quelli previsti sono concordanti: i coefficienti di correlazione tra le due serie di valori sono 0,994 per l'angolo di avvitamento dell'elica *B*, 0,917 per l'angolo di oscillazione dell'elica *B*, 0,991 per l'angolo di avvitamento elicoidale dell'ottamero *A* e 0,995 per l'angolo di oscillazione dello pseudo-ottamero *A*. Quest'ultimo (che è costituito in realtà da due tetrameri sovrapposti, tenuti uniti da fosfati) si comporta sotto questo aspetto come un vero ottamero. La sovrapposizione delle coppie di basi sembra essere più importante di quanto l'impalcatura costituita dai gruppi fosfato lo sia nel connettere la struttura a elica locale.

due lati di un valore pari alla metà. Pertanto il contributo complessivo alla funzione sommatoria per ciascuna alternanza purina-pirimidina e per le due alternanze che la fiancheggiano sarà + 1, – 2, + 1, mentre i valori per ogni alternanza pirimidina-purina e per le due alternanze che la fiancheggiano saranno + 2, – 4, + 2.

3. Si aggiungano i valori relativi a tutte le coppie e si ottenga così la funzione sommatoria completa Σ_1, la quale misura la variazione locale, indotta dalla sequenza di basi, partendo dall'angolo medio di avvitamento a elica.

La derivazione della sommatoria Σ_1 viene illustrata nel caso del dodecamero *CGCGAATTCGCG* dell'elica *B* nella figura della pagina a fronte. Sono parimenti confrontati i valori previsti e quelli osservati. Un confronto mediante regressione lineare degli angoli osservati con la sommatoria Σ_1 dà un coefficiente di correlazione di 0,994, correlazione estremamente significativa tra le due quantità. La funzione sommatoria riproduce fedelmente gli allargamenti e le riduzioni degli angoli in osservazione tranne alle due estremità della sequenza, dove la sovrapposizione intermolecolare nel cristallo distorce l'elica.

Un analogo procedimento può essere seguito per calcolare la sommatoria Σ_2, che misura la variazione locale nell'angolo di oscillazione, ma le correzioni in corrispondenza delle alternanze pirimidina-purina hanno segno inverso (–2, +4, –2) dato che, in corrispondenza di esse, l'impedimento sterico è ridotto dall'allargamento dell'angolo sul lato del solco minore, il che rende tale angolo ancora più positivo. La derivazione della Σ_2 e il suo confronto con gli angoli di oscillazione che sono stati osservati vengono illustrati per il dodecamero *CGCGAATTCGCG*. Anche qui l'allargamento e la riduzione di tali angoli possono essere previsti con precisione tranne alle estremità dell'elica.

Se queste due funzioni sommatorie valessero solo per un'elica di un determinato tipo, il metodo assomiglierebbe a molte brillanti idee scientifiche che funzionano una volta nelle mani del loro inventore, ma mai di nuovo nelle mani di qualcun altro. Risulta, tuttavia, che quando le stesse regole vengono applicate alle quattro eliche *A* note, i risultati che si ottengono sono altrettanto buoni di quelli ottenuti per la sequenza *B*. Una delle strutture *A* ha fornito una prova particolarmente convincente. Le due molecole tetramere di *CCGG*/ /*CCGG* sono impaccate l'una sopra l'altra nel cristallo e formano un ottamero anche se le due coppie di basi al centro non sono unite da fosfati. Malgrado questa assenza di connessione covalente, sia l'angolo di avvitamento elicoidale sia l'angolo di oscillazione del piano delle basi in corrispondenza dell'alternanza centrale sono esattamente quelli che si sarebbero potuti prevedere dalle funzioni sommatorie derivate per un ottamero covalente, intatto. Risulta così che le forze di impaccamento tra coppie di basi prevalgono sull'influenza esercitata dall'impalcatura zucchero-fosfato nella determinazione della struttura a doppia elica del DNA. Sia la funzione che si riferisce all'avvitamento sia quella che si riferisce all'oscillazione valgono anche per l'ottamero a elica *A GGTATACC*; soltanto la funzione relativa all'avvitamento comincia a non andare più bene nel caso delle coppie di basi dell'ibrido RNA-DNA nell'elica mista (*GCG*)*TATACGC*.

Dalla struttura al controllo

Queste fluttuazioni locali nella struttura a elica influiscono davvero sulla lettura delle sequenze di basi del DNA da parte dei repressori e di analoghe proteine di controllo? Se si accetta l'idea che il processo di decifrazione comporti la formazione di legami a idrogeno tra le catene laterali proteiche e gli atomi di azoto e di ossigeno sui margini delle basi, è difficile immaginare che queste variazioni di anche 15 gradi dalla media non abbiano un effetto importante sul processo di riconoscimento. Forse le fluttuazioni indotte dalla sequenza delle basi rappresentano una fine sintonizzazione dell'adattamento e dell'incastro tra DNA e proteina, aggiungendo al processo di decifrazione una dimensione che va oltre la semplice dislocazione (o distribuzione) di donatori e accettori di legami a idrogeno lungo il margine di ogni coppia di basi.

Se le cose stiano proprio in questi termini può essere stabilito soltanto dall'analisi della struttura cristallina, mediante raggi X, effettuata per una proteina di riconoscimento, legata in un complesso con la propria particolare sequenza di DNA. Parecchi studi su complessi di questo genere sono ora in corso e interessano i repressori *lac*, *lambda* e *cro*, la proteina attivatore di cataboliti (CAP) e vari enzimi di restrizione. John Rosenberg dell'Università di Pittsburgh ha costruito di recente una mappa di densità elettronica per l'enzima di restrizione *Eco R*I, legato al proprio sito di riconoscimento sul dodecamero *CGCGAATTCGCG*. Secondo lui, il DNA legato rimane nella forma *B*, ma l'elica viene distorta.

Gli interrogativi che ci si può sensatamente porre sono sempre influenzati da ciò che già si conosce. Un nuovo passo avanti nella conoscenza può togliere un argomento dal regno dell'oziosa speculazione per porlo al centro dell'indagine scientifica. Questo è ciò che le recenti analisi ai raggi X, condotte su cristalli singoli, hanno fatto per domande quali: in che modo la sequenza delle basi nella doppia elica del DNA influisce sulla struttura e sul comportamento dell'elica stessa? In quale misura questo effetto è significativo nella decifrazione dell'informazione? Le risposte cominciano a emergere solo ora, ma vale bene la pena che le domande siano poste: esse finiranno per condurre, infatti, a una miglior comprensione del controllo genetico.

L'aerodinamica dei veicoli a propulsione umana

Anche se una bicicletta e chi le sta in sella sono ostacolati dalla resistenza opposta dall'aria, soluzioni aerodinamiche consentono velocità di quasi 100 chilometri all'ora su strada pianeggiante

di Albert C. Gross, Chester R. Kyle e Douglas J. Malewicki

Da decenni ormai i principi dell'aerodinamica vengono applicati con grande successo al miglioramento della velocità e del rendimento di aeroplani, automobili, motociclette e perfino di sciatori e pattinatori impegnati a livello agonistico. Fino a pochissimo tempo fa però i veicoli la cui forza motrice è costituita dall'energia muscolare umana sono rimasti praticamente ignorati, e la cosa è strana, se si tien conto del fatto che la resistenza aerodinamica è la forza frenante di gran lunga più importante che agisca su di essi. Nel caso delle biciclette, per esempio, essa è responsabile di oltre l'80 per cento della forza totale che entra in gioco, rallentando quelle che procedono a velocità superiori ai 28 chilometri all'ora. Qui cercheremo di spiegare questa trascuratezza e di mostrare ciò che l'attenzione prestata all'aerodinamica sta incominciando a fare per le prestazioni dei veicoli terrestri mossi dall'energia muscolare umana.

Considerando innanzitutto la bicicletta, si vede subito che da quasi un secolo a questa parte la sua forma è rimasta pressoché immutata. Il Rover Safety Cycle, che fu introdotto in Inghilterra nel 1884, potrebbe facilmente passare per una bicicletta moderna. Gli mancano soltanto la struttura di sostegno della sella, che avrebbe poi costituito il moderno telaio a losanga, e alcuni componenti come i freni e la moltiplica. L'importanza dell'aerodinamica era stata riconosciuta fin quasi dal principio sia da chi progettava le biciclette sia da chi se ne serviva, ma costrizioni artificiose imposte al disegno hanno impedito in larga misura l'applicazione della necessaria tecnologia. Era ovvio allora, così come lo è ai nostri giorni, che alla velocità di una bicicletta da corsa, variabile da 30 a 48 chilometri all'ora, le forze dell'aria sono enormi.

Prima del 1900 la posizione curva sul manubrio del corridore ciclista era diventata comune come modo per ridurre la resistenza dell'aria. Un'altra pratica adottata prima del 1900 era quella di proteggere dal vento un ciclista facendolo precedere da una bicicletta a posti multipli. Nel 1895 il ciclista gallese Jimmy Michael percorse in un'ora 28,6 miglia (45,76 chilometri) dietro una bicicletta a quattro posti. Nel 1899 l'americano Charles Murphy, soprannominato «Miglio al minuto», acquistò fama internazionale percorrendo un miglio a 63,24 miglia (101,18 chilometri) all'ora su una bicicletta che procedeva dietro un treno della Long Island Rail Road su una «pista» di tavole di legno costruita per l'occasione.

Nel 1912 il francese Étienne Bunau-Varilla brevettò una struttura aerodinamica che, ispirandosi alla forma dei primi dirigibili, racchiudeva tanto la bicicletta quanto il ciclista. Versioni diverse di questa bicicletta e i suoi discendenti stabilirono in Europa dal 1912 al 1933 vari primati di velocità. Nel 1933 il francese Marcel Berthet coprì in un'ora 49,69 chilometri in sella a un mezzo a profilo aerodinamico detto Vélodyne; la sua andatura era di 4,8 chilometri all'ora superiore al primato dell'epoca su una bicicletta standard.

Nello stesso anno l'inventore francese Charles Mochet costruì una bicicletta *recumbent* (sulla quale il ciclista pedalava stando steso sul dorso), alla quale in seguito diede una forma più aerodinamica. Guidato da François Faure, un corridore professionista, questo «Vélocar» stabilì tra il 1933 e il 1938 numerosi primati di velocità. Mochet e Faure speravano che tali primati venissero riconosciuti dall'Union Cycliste Internationale, l'organo direttivo del ciclismo mondiale, ma le loro speranze andarono deluse.

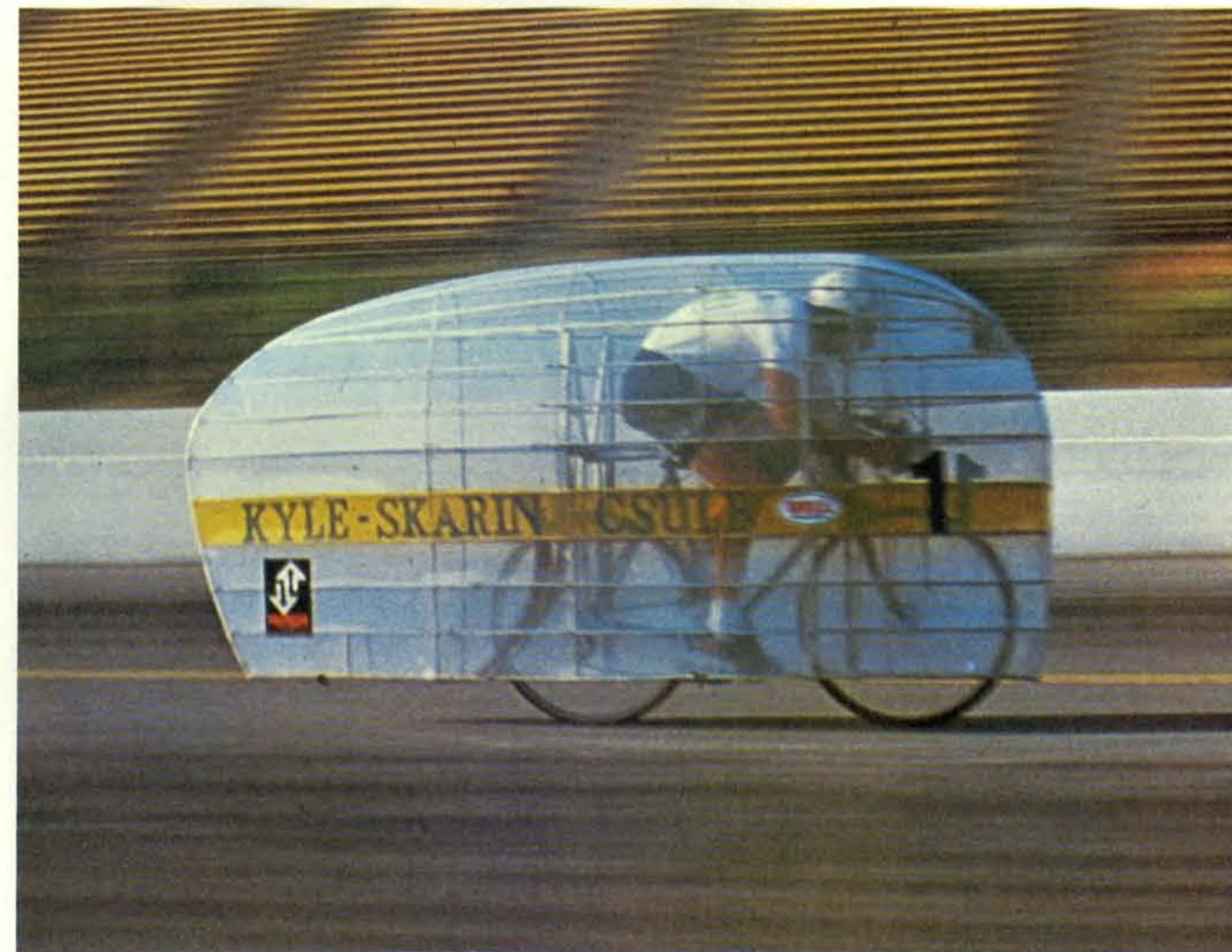
Al contrario, nel 1938 l'Union proibì nelle gare ufficiali l'uso di congegni aerodinamici e di biciclette *recumbent*, e la norma è tuttora in vigore. Questa proibizione ha rappresentato un grosso ostacolo allo sviluppo di biciclette ultraveloci ed è una delle due ragioni principali per le quali la bicicletta è rimasta pressoché immutata per così tanto tempo. (L'altra ragione è costituita dal fatto che nei paesi sviluppati il passaggio all'automobile ha reso la bicicletta meno importante di un tempo come mezzo di trasporto.)

Con la sua decisione l'Union classificava praticamente come un «imbroglio» qualsiasi miglioramento apportato all'aerodinamica della bicicletta e qualsiasi altro mutamento tecnologico. (È una fortuna forse che l'Union non esistesse ancora quando un veterinario scoto-irlandese, John Boyd Dunlop, creò nel 1887 gli pneumatici, altrimenti, chissà, continueremmo forse ad andare su biciclette e su automobili con ruote di acciaio piene.) A poco a poco però, e ciò va a suo credito, l'Union ha incominciato ad allentare i freni per quel che riguarda i mutamenti in fatto di aerodinamica, anche se le *recumbent* sono ancora proibite. Dal 1976 sono diventate comuni nelle gare ciclistiche internazionali le tute intere molto aderenti. Sono stati permessi anche i caschi a profilo aerodinamico, le sezioni trasversali a goccia per i tubi del telaio, le leve dei freni aerodinamiche e miglioramenti vari, sul piano aerodinamico, di altre componenti. In effetti, in ogni tipo di veicolo mosso dalla forza dell'uomo i mutamenti tecnologici si susseguono a un ritmo senza riscontro dai tempi d'oro della bicicletta nel secolo scorso.

Questi rapidi mutamenti si possono attribuire in parte a una serie di fatti avvenuti in California. Nel 1973 furono costruite e sperimentate da uno di noi (Kyle) e da Jack H. Lambie, consulente di aerodinamica che lavorava per conto proprio, le prime due biciclette a profilo aerodinamico degli Stati Uniti. A differenza dei loro predecessori, Kyle e Lambie calcolarono effettivamente la riduzione della resistenza dell'aria che si può ottenere dando a un veicolo una forma aerodinamica. E la calcolarono effettuando numerosi esperimenti a ruota libera, in cui un veicolo privo di forza motrice veniva fatto decelerare su una superficie piana. In queste condizioni, la decelerazione del veicolo è proporzionale al totale delle forze frenanti che agiscono su di esso; la velocità e la decelerazione vengono misurate da appositi strumenti. Pubblicando ognuno per conto proprio i loro risultati, Kyle e Lambie conclusero entrambi che, con una carenatura verticale di forma alare che racchiudesse completamente tanto la bicicletta quanto il ciclista, era possibile ridurre di oltre il 60 per cento le forze frenanti. (Solo un paio di anni dopo Kyle e Lambie vennero a sapere che veicoli del genere erano stati costruiti già da tempo in Europa.)

Nel 1974 Ronald P. Skarin, un ciclista olimpico statunitense, stabilì alla Naval Air Station di Los Alamitos cinque primati di velocità in sella alla bicicletta a profilo aerodinamico di Kyle. Quest'ultimo e Lambie decisero di organizzare una corsa per veicoli a propulsione umana senza limitazioni di sorta. A questa storica prima corsa, che si svolse il 5 aprile 1975 a Irwindale, in California, parteciparono 14 veicoli peculiari, molti dei quali erano delle biciclette *recumbent*, su cui in alcuni casi il guidatore pedalava supino (faccia in su) e in altri prono (faccia in giù). Alcuni veicoli erano spinti sia a forza di piedi sia a forza di mani. Il vincitore, alla media di 44,87 miglia (71,79 chilometri) all'ora, fu un tandem a profilo aerodinamico progettato da Philip Norton, un insegnante di scuola superiore vicino a Claremont, in California. I pedalatori erano Norton e Christopher Deaton, che è un bravo ciclista, ma non un corridore di classe mondiale. (La bicicletta da corsa standard più veloce e priva di qualsiasi mezzo ausiliario che sia mai stata usata ha toccato la velocità di 69,52 chilometri all'ora, un primato stabilito nel 1982 dal sovietico Sergei Kopylov, un ciclista di levatura mondiale.)

Di fronte alla politica dell'Union Cycliste Internationale, avversa alle forme aerodinamiche, i concorrenti di questa gara fondarono nel 1976 la International Human Powered Vehicle Association, il cui scopo era convalidare le gare alle quali i veicoli a propulsione umana non fossero soggetti a limitazioni progettuali di sorta. Da allora in decine di gare svoltesi in molti paesi le macchine sono diventate molto più raffinate e la velocità è aumentata costantemente. Quattro veicoli hanno superato il limite delle 55 miglia (88 chilometri) all'ora fissato negli Stati Uniti per le automobili. (Ognuno ha ricevuto dalla California Highway Patrol una multa onoraria per eccesso di velocità.) Fra questi c'è un quadriciclo a profilo aerodinamico della terza generazione progettato da Norton.

Attualmente il veicolo a propulsione umana più veloce del mondo è il Vector Tandem, un veicolo *recumbent* per due persone piacevolmente aerodinamico, costruito da un gruppo guidato da Allan A. Voigt, un ingegnere che, quale presidente della Versatron Research Inc., progetta soprattutto servomotori aerospaziali. (I pedalatori sono in posizione supina, con il viso rivolto in versi opposti.) Nel

La bicicletta aerodinamica da corsa progettata da uno degli autori (Kyle) è ripresa mentre, guidata da Ronald P. Skarin, un ciclista olimpico americano, stabilisce, con 31,88 miglia (50,88 chilometri), il primato mondiale dell'ora con partenza da fermo. La prestazione è stata resa possibile dalla carenatura aerodinamica, che riduce la resistenza opposta dall'aria al ciclista e alla bicicletta. Skarin stabilì il nuovo primato di velocità nel 1979 all'Ontario Motor Speedway di Ontario, in California. A parte la carenatura, il veicolo era una comune bicicletta da corsa.

1980, con una partenza lanciata di circa un chilometro e mezzo, il veicolo percorse 200 metri sulla carreggiata dell'Ontario Motor Speedway in California alla velocità di 62,92 miglia (100,67 chilometri) all'ora. Più tardi quello stesso anno il Vector Tandem percorse 40 miglia (64 chilometri) sulla Interstate Route 5 tra Stockton e Sacramento alla media di 50,5 miglia (80,80 chilometri) all'ora.

Queste velocità straordinarie dipendono in buona misura dall'attenzione rivolta all'aerodinamica. Un ciclista che proceda a 32 chilometri all'ora sposta circa 500 chilogrammi di aria al minuto. Quando non hanno nulla di aerodinamico, la macchina e l'uomo lasciano una scia notevole ed esigono un pesante tributo in fatto di energia muscolare.

Due tipi di resistenza aerodinamica incidono sulle prestazioni di una bicicletta: la resistenza di pressione (o di forma) e la resistenza d'attrito. La prima si ha quando il flusso dell'aria non segue i contorni del corpo che si muove. Questa separazione modifica la distribuzione della pressione dell'aria sul corpo stesso. Se la separazione avviene verso la parte posteriore del corpo, lì la pressione dell'aria diventa inferiore a quella che si registra sulla superficie anteriore, causando resistenza.

La resistenza d'attrito è dovuta alla viscosità dell'aria, prodotta dalle forze di taglio generate nello strato limite, vale a dire nello strato d'aria immediatamente adiacente alla superficie del corpo.

Certe configurazioni smussate che si trovano su una bicicletta, come cilindri, sfere e altre forme, sono aerodinamicamente inefficienti perché il flusso dell'aria si separa dalla loro superficie. Dietro gli oggetti si formano quindi zone di pressione più ridotta, che danno vita a una resistenza di pressione centinaia di volte superiore a quella di attrito. Intorno a una forma aerodinamica, per contro, l'aria fluisce con regolarità, chiudendosi dietro al corpo via via che quest'ultimo procede. La resistenza di pressione si riduce così in misura notevole mentre diventa più importante quella di attrito.

Per il massimo rendimento un veicolo dovrebbe essere progettato in modo da ridurre al minimo il trasferimento all'aria, da parte dei due tipi di resistenza, di energia irrecuperabile. Allo stato attuale della tecnologia, la resistenza aerodinamica assorbe dal 40 al 50 per cento dell'energia da combustibile consumata da un'automobile o da un autocarro che procedano alla velocità di 90 chilometri all'ora. Dal momento che, rispetto alle automobili e agli autocarri, la bicicletta ha un peso, una potenza e una resistenza al rotolamento inferiori, nonché una scarsa aerodinamicità, a velocità superiori ai 16 chilometri all'ora la resistenza aerodinamica rappresenta una percentuale ancora maggiore dell'energia consumata.

Un termine usato per definire l'efficienza aerodinamica di una forma è il coefficiente di resistenza. Una forma inefficiente come una sfera avrà, poniamo, un coefficiente di resistenza di 1,3, mentre una forma aerodinamica come una goccia ne avrà uno inferiore a 0,01. Perciò un oggetto di forma a goccia potrà muoversi con una perdita di energia pari a meno di un decimo di quella di un oggetto di forma cilindrica.

Per i veicoli di trasporto su strada la resistenza aerodinamica è quasi direttamente proporzionale al prodotto della superficie frontale per il coefficiente di resistenza. Per comodità diamo a questo prodotto il nome di superficie frontale effettiva. Per stabilire quale di due veicoli abbia una resistenza aerodinamica più ridotta non basta confrontare i rispettivi coefficienti di resistenza; bisogna tener conto anche delle dimensioni del veicolo. È quel che si fa nel concetto di superficie frontale effettiva. Una normale bicicletta e chi le sta in sella hanno una superficie frontale effettiva che va da 3150 a 5500 centimetri quadrati mentre in un veicolo aerodinamico a motore umano tale superficie può essere inferiore a 460 centimetri quadrati.

La forza della resistenza aerodinamica aumenta in proporzione al quadrato della velocità. Siccome la potenza è proporzionale al prodotto della forza di resistenza per la velocità, la potenza necessaria per spingere attraverso l'aria un oggetto aumenta in proporzione al cubo della velocità. Un modesto aumento di velocità richiede pertanto un enorme aumento di potenza. Un ciclista che raddoppi all'improvviso la propria produzione di potenza mentre sta procedendo a 32 chilometri all'ora aumenterà di poco la propria velocità, portandola soltanto a 41,6 chilometri all'ora.

Per contro, un'eventuale riduzione della resistenza aerodinamica incide sulla velocità meno di quanto si possa pensare. Se si riduce della metà la resistenza che l'aria oppone a un veicolo che procede a 32 chilometri all'ora, un ciclista che non cambi la propria produzione di potenza accelererà soltanto fino a 39,04 chilometri all'ora. La ragione va ricercata nel fatto che la resistenza al rotolamento rimane costante. Se fosse possibile ignorare questa resistenza, basterebbe raddoppiare i watt o ridurre della metà la superficie frontale effettiva per riportare la velocità a circa 41,6 chilometri all'ora.

Per riassumere: le alte velocità richiedono un'efficienza aerodinamica estremamente elevata. Con un input di circa 750 watt da parte di ognuno dei suoi due ciclisti, il Vector Tandem raggiunse una velocità di 100,67 chilometri all'ora. Per raggiungere questa velocità una bicicletta standard avrebbe bisogno di oltre 4500 watt, una potenza chiaramente al di fuori delle possibilità di un essere umano.

Nel caso dei veicoli a propulsione umana, progettisti e ciclisti possono ridurre la resistenza aerodinamica soprattutto in tre modi. In primo luogo, possono ridurre la quantità di energia sprecata nell'interazione del veicolo con l'aria. A tale scopo si dà un profilo aerodinamico (cambiandone la forma) alla parte anteriore e a quella posteriore degli oggetti ottusi, in modo da ridurre al minimo la resistenza di pressione, e si appianano le superfici irregolari, in modo da ridurre al minimo la resistenza di attrito. In secondo luogo, si può ridurre la quantità di aria che si incontra ogni secondo quando si procede su qualsiasi percorso. Ciò si ottiene riducendo la superficie frontale effettiva della combinazione veicolo-conducente. Lo stesso effetto si può ottenere pedalando ad altitudini elevate. In terzo luogo, il ciclista può trovare dell'aria che si muova in modo tale da fornire un vento di coda, ossia un vento a favore. Qui l'impostazione più efficace è quella, come si dice in gergo, di farsi «tirare», di pedalare cioè nella scia di un altro veicolo a distanza molto ravvicinata.

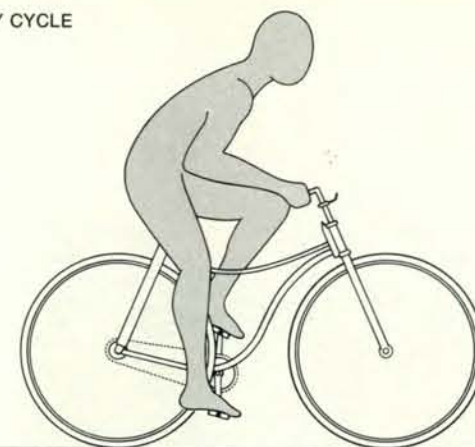
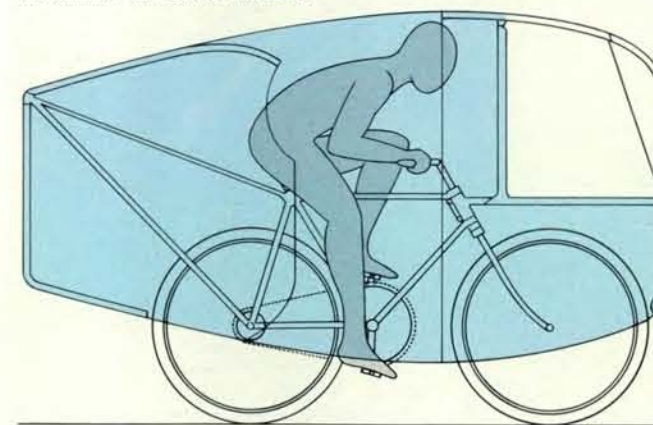
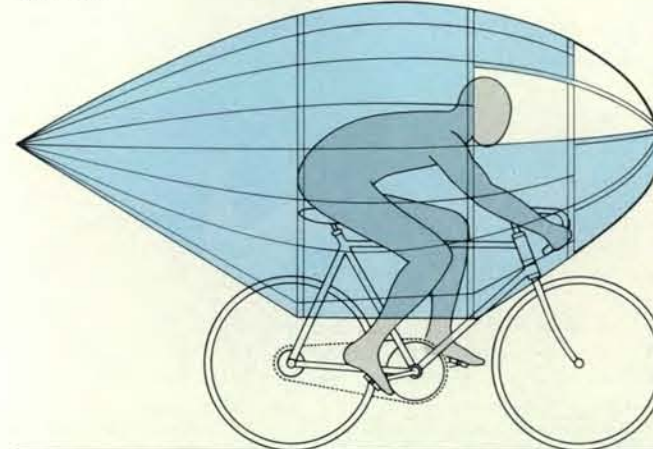
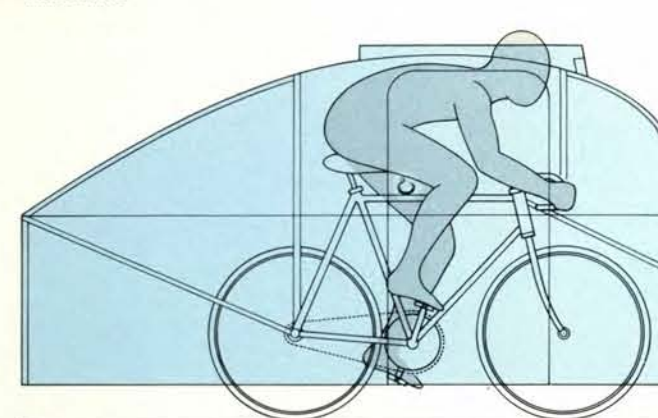
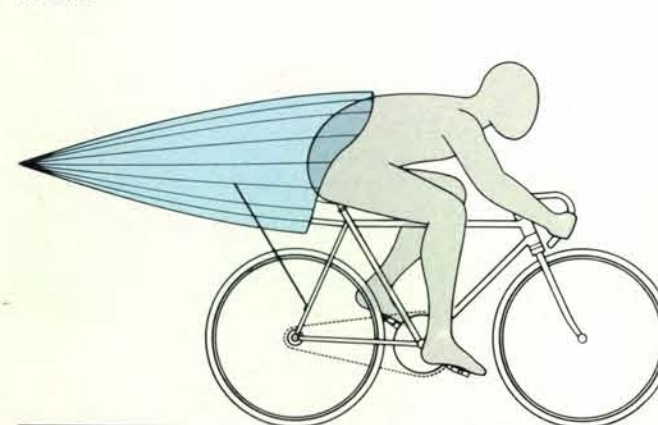
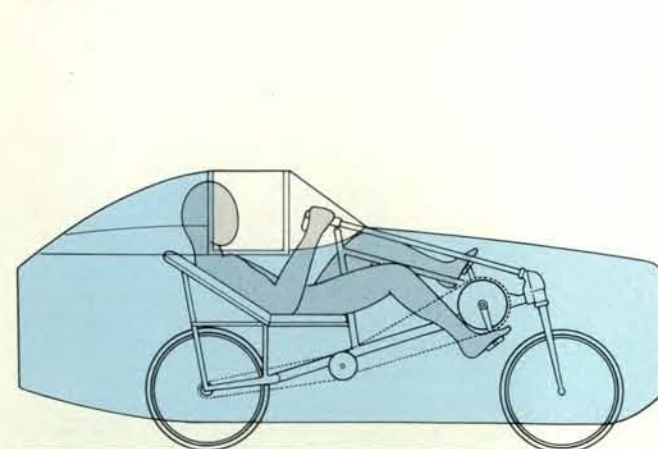
Ad altitudini elevate l'atmosfera è meno densa e le biciclette incontrano meno aria. A Città del Messico (altitudine 2277 metri, dove l'aria ha una densità pari soltanto all'80 per cento di quella al livello del mare) i primati ciclistici sono dal 3 al 5 per cento più elevati di quelli stabiliti ad altitudini inferiori. A La Paz, in Bolivia (altitudine 3630 metri), sarebbe possibile in teoria migliorare del 14 per cento i primati stabiliti a livello del mare. Sulla Luna, dove non c'è atmosfera e dove l'attrazione gravitazionale è pari solo a un sesto di quella terrestre, un ciclista opportunamente equipaggiato potrebbe in teoria pedalare a 380 chilometri all'ora con un input molto modesto di 75 watt.

Analizzando la relazione, secondo la quale l'80 per cento della potenza generata da un ciclista che proceda su un terreno pianeggiante alla velocità di circa 29 chilometri all'ora se ne va per vincere la resistenza dell'aria, si scopre che circa il 70 per cento del consumo di potenza è dovuto alla resistenza opposta dall'aria al ciclista e il 30 per cento alla resistenza opposta dall'aria alla bicicletta. Questa scoperta porta a concludere che, per migliorare le prestazioni di una bicicletta standard, è necessario innanzi tutto migliorare l'aerodinamica del ciclista.

Per i ciclisti che partecipano alle gare, le limitazioni dell'Union Cycliste Internationale non lasciano molto spazio ai miglioramenti oltre a ciò che è stato già fatto adottando la posizione curva sul manubrio, il casco aerodinamico, la tuta molto aderente e la forma aerodinamica di alcuni componenti della bicicletta. Come ha calcolato Voigt, anche con una bicicletta «perfetta» (nessuna resistenza aerodinamica esercitata sulla macchina a qualsiasi velocità e pneumatici senza resistenza al rotolamento), la sola resistenza opposta dall'aria al ciclista impedirebbe seriamente qualsiasi miglioramento delle prestazioni. Secondo Voigt, un ciclista curvo sul manubrio di una bicicletta da corsa convenzionale potrebbe raggiungere una velocità massima di circa 54 chilometri all'ora con un input di potenza di 750 watt. Su una bicicletta perfetta lo stesso ciclista che producesse lo stesso sforzo potrebbe raggiungere i 61 chilometri all'ora.

Per i milioni di ciclisti non agonistici che vogliono semplicemente un'andatura più efficiente sono possibili parecchi miglioramenti aerodinamici, che si possono classificare in ordine di costo, partendo da quello più economico: una carenatura parziale come lo Zzipper, sviluppato e fabbricato da Glen Brown di Santa Cruz, in California. Si tratta di un piccolo schermo trasparente e aerodinamico montato davanti al guidatore. Per circa 60 dollari un ciclista può così ridurre di circa il 20 per cento la resistenza aerodinamica, ottenendo un incremento di velocità di circa quattro chilometri all'ora per un input di 750 watt.

Un altro modo efficace per ridurre la resistenza aerodinamica è quello di inforcare una bicicletta in cui il ciclista è in posizione distesa. (Il veicolo verrebbe a costare parecchie centinaia di migliaia di lire più di una normale bicicletta da turismo.) I pionieri in questo campo sono Gardner Martin di Freedom, in California, progettista dell'Easy Racer, e David Gordon Wilson del Massachusetts Institute of Technology, progettista dell'Avatar 2000. Data la superficie frontale più ridotta presentata dal ciclista in posizione distesa, la resistenza dell'aria diminuisce del 15-20 per cento, producendo più o meno lo stesso aumento di velocità ottenuto con la carenatura Zzipper.

La bicicletta *recumbent* offre comun-

I primi miglioramenti apportati ai veicoli su strada a propulsione umana ebbero come risultato l'introduzione in Inghilterra, nel 1884, del Rover Safety Cycle. Nel 1912 e nel 1913 il francese Étienne Bunau-Varilla ottenne brevetti per un progetto aerodinamico; biciclette analoghe stabilirono molti primati di velocità. La Goricke fu sviluppata in Germania nel 1914. Nel 1933 la Vélodyne, guidata dal francese Marcel Berthet, percorse in un'ora 49,69 chilometri. Dello stesso anno è il Rocket, progettato da Oscar Egg. Il Vélocar stabilì tra il 1933 e il 1938 parecchi primati di velocità. Quasi tutti i disegni si basano su dati tratti dal Wolfgang Gronen Archiv di Binningen, in Germania.

que altri vantaggi. È più comoda da montare di una bicicletta standard. Negli incidenti che non comportano lo scontro con un'automobile è molto meno pericolosa, in quanto il ciclista è più vicino al suolo (rendendo meno gravi le cadute) e i piedi sono avanti (rendendo meno probabili in una caduta le lesioni al capo). C'è, sì, il problema che sulla strada una bicicletta di questo genere è difficile da scorgere ed è forse quindi più vulnerabile nei confronti delle automobili, ma si tratta di un problema che può essere in parte risolto montando sul veicolo un'asta lunga e sottile con una bandiera.

La soluzione più costosa è una bicicletta dotata di una carenatura completa. La Vector Single, una versione monoposto del Vector Tandem, è il migliore esempio di un veicolo a pedali interamente carenato, chiuso. (È la macchina presentata sulla copertina di questo fascicolo.) Secondo Voigt, il veicolo è in grado in linea teorica di toccare i 98,72 chilometri all'ora con un input di 750 watt, un incremento di 45,12 chilometri all'ora rispetto alle prestazioni fin qui ottenute con una bicicletta da corsa standard. Una Vector Single costa grosso modo quanto una bicicletta da corsa di prim'ordine.

Anche in salita o in discesa un veicolo completamente aerodinamico conserva il proprio vantaggio su una bicicletta convenzionale. Pur pesando circa 36 chilogrammi, rispetto ai poco più di 11 di una bicicletta standard, la Vector Single può superare salite moderate alla stessa velocità o più velocemente della bicicletta. Con un input di 300 watt, una bicicletta può superare una pendenza del 2,5 e una del 6 per cento alla velocità rispettivamente di circa 25,6 e 17,6 chilometri all'ora. Con lo stesso input la Vector può superare le due pendenze rispettivamente a 32,8 e 17,6 chilometri all'ora.

In discesa la differenza fra le due macchine è notevole. La bicicletta può percorrere una pendenza del 2,5 per cento a 46,4 chilometri all'ora, la Vector a 86,4. Su una pendenza del 6 per cento la bicicletta può raggiungere una velocità di 62,4 chilometri all'ora, mentre la Vector può superare i 160. Tali velocità potenziali stanno a indicare che, se i veicoli aerodinamici a propulsione umana diventeranno comuni, bisognerà prestare particolare attenzione al progetto dei freni e della sospensione e alla stabilità del veicolo.

Poiché la resistenza aerodinamica è proporzionale al quadrato della velocità relativa, i venti di prua, i venti di coda e perfino i venti di traverso possono modificare drasticamente sia la resistenza aerodinamica sia il fabbisogno di potenza. Per esempio, un ciclista che proceda a circa 29 chilometri all'ora in aria calma deve accrescere del 100 per cento la propria produzione di potenza per mantenere quella velocità contro un vento di prua di 16 chilometri all'ora. Di solito un ciclista che incontri un vento di prua rallenta e cerca di mantenere la spinta abituale delle gambe e la cadenza della pedalata azionando il cambio di velocità. Questa è una delle ragioni per le quali le biciclette dotate di («più rapporti») moltiplica sono preferibili anche su percorsi pianeggianti.

Un vento di coda fa procedere più velocemente il ciclista mantenendo il normale input di potenza. In genere l'aria in movimento accelera o rallenta una bicicletta in misura pari grosso modo alla metà della velocità del vento. Quando una bicicletta procede nella scia di un'altra, il fabbisogno di potenza del ciclista che sta dietro si riduce di circa il 30 per cento, in quanto il ciclista che sta davanti, quello che «tira», crea un vento di coda artificiale.

Quanto più da vicino una bicicletta segue quella che «tira», tanto più marcato è l'effetto della trazione. Si può pensare che il ciclista che sta dietro di un tandem si faccia «tirare» da distanza estremamente ravvicinata. Le due persone in sella a un tandem usano ciascuna il 20 per cento in meno di potenza di due ciclisti separati.

Quando ciclisti in fila si danno il cambio nella posizione di testa, tutto il gruppo può andare a una velocità molto superiore a quella di un ciclista isolato. In una gara a inseguimento sui 4000 metri, una

ZZIPPER

VECTOR SINGLE

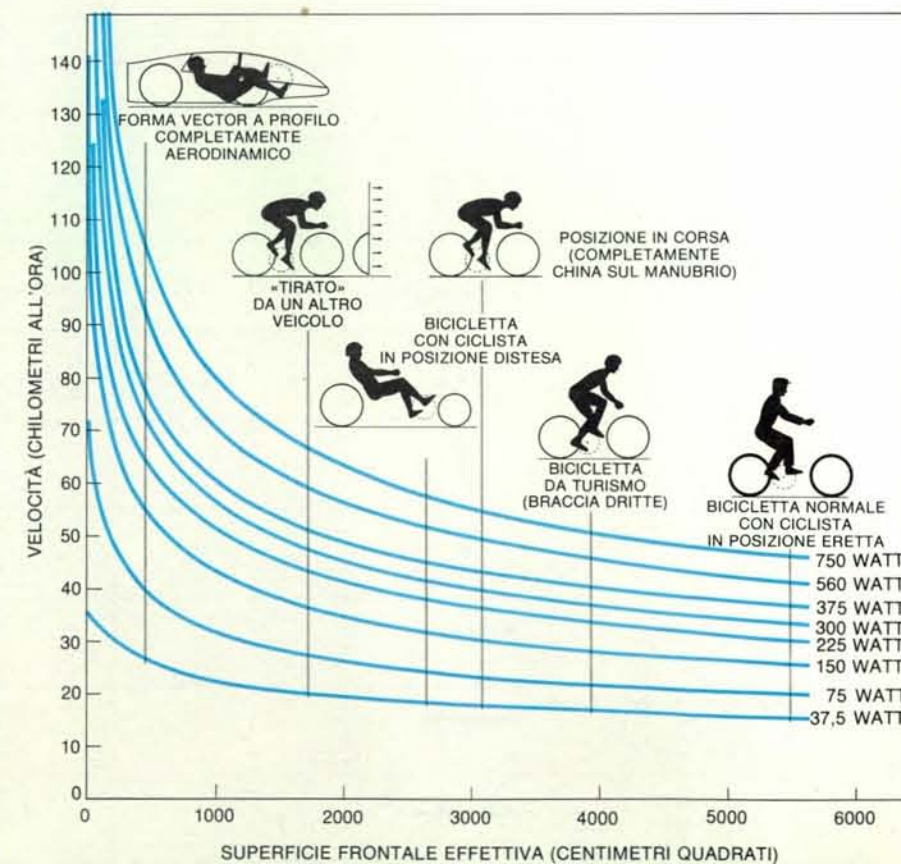
KYLE STREAMLINER

EASY RACER

AVATAR 2000

VEICOLO SCHÖNDORF OGNI-TEMPO

I moderni veicoli a propulsione umana fanno uso intenso del profilo aerodinamico per ridurre la resistenza opposta dall'aria alla combinazione veicolo-uomo. Il più semplice è lo Zzipper, una carenatura parziale montata davanti al ciclista. Il Kyle Streamliner risale al 1973. Un progetto rivolto più ai turisti e ai pendolari che all'agonismo è l'Avatar 2000, che sfrutta per i ciclisti i vantaggi della posizione distesa. Il Vector Single, che ha una carenatura completa, è teoricamente in grado di raggiungere quasi i 100 chilometri all'ora con un input, da parte del ciclista, di 750 watt. L'Easy Racer è una bicicletta *recumbent* per turisti e pendolari, ma è stata usata anche nelle corse. L'ultimo veicolo è una delle *recumbent* utilizzabili in qualsiasi condizione di tempo, progettate dal tedesco Paul Schöndorf per persone anziane e per handicappati.

I primati di velocità ottenuti con veicoli terrestri a propulsione umana sono aumentati rapidamente dopo la fondazione nel 1976 dell'International Human Powered Vehicle Association, che non pone limiti progettuali di sorta per la partecipazione alle gare. L'anno della fondazione è indicato dalla linea tratteggiata. Per molti anni prima di allora le regole dell'Union Cycliste Internationale, che bandivano dalle gare ciclistiche ufficiali i veicoli a profilo aerodinamico, avevano mantenuto praticamente immutati i primati di velocità. Le curve rappresentano i primati per più corridori sui 200 metri con partenza lanciata (*in colore*), per corridori singoli nelle stesse condizioni (*in grigio*) e per corridori che pedalavano per un'ora producendo il massimo sforzo (*in nero*).

Un profilo aerodinamico ha l'effetto di migliorare le prestazioni dei veicoli a propulsione umana a tutti i livelli di input di potenza. La bicicletta da turismo guidata da un ciclista in posizione eretta è il veicolo meno aerodinamico, mentre il modello Vector è il più aerodinamico. Farsi «tirare» significa seguire da vicino un altro veicolo, in questo caso una bicicletta. Un buon atleta e un non atleta sano possono produrre 750 watt rispettivamente per circa 30 secondi e 12 secondi e sono in grado di continuare a produrre rispettivamente 300 e 75 watt per circa otto ore. La superficie frontale effettiva è data dal coefficiente di resistenza moltiplicato per la superficie frontale prevista.

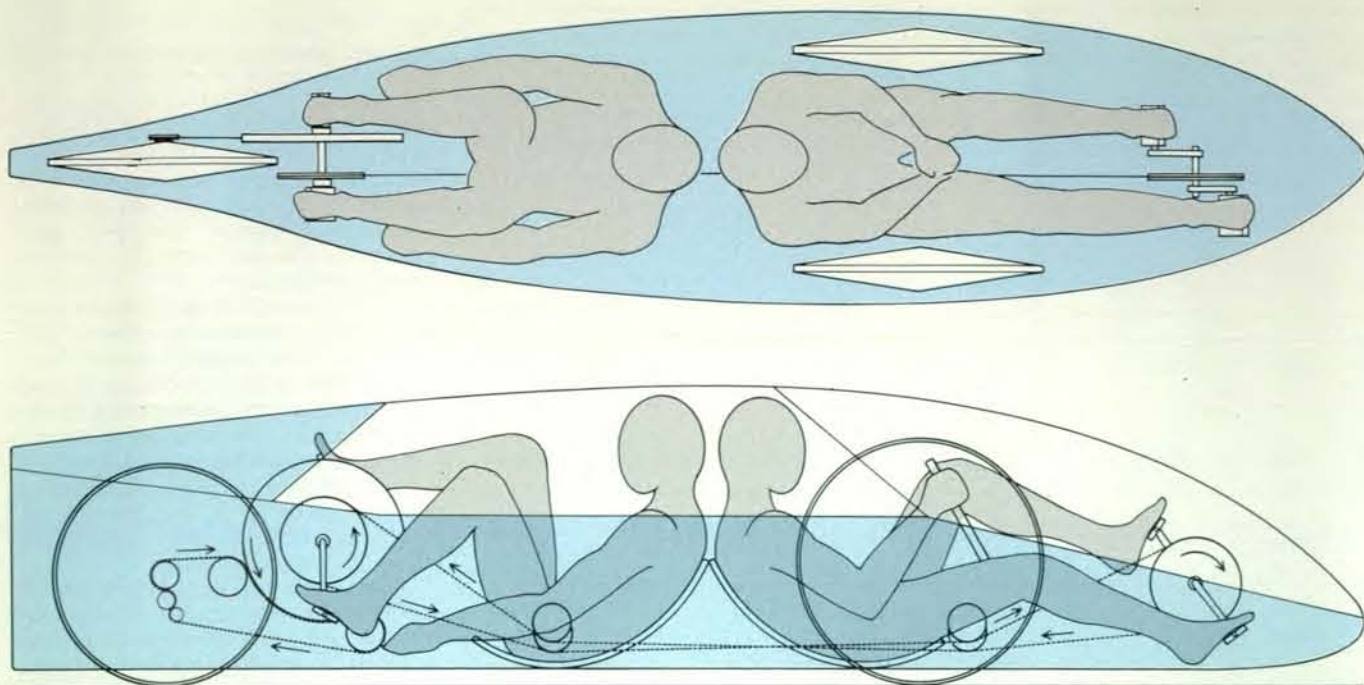
	DESCRIZIONE VEICOLO	PESO DEL VEICOLO (kg)	PESO DEL CICLISTA (kg)	PNEUMATICI			FORZE A 32 km/h (kg)	COEFF. DI RESISTENZA AL ROTOLAMENTO	DATI AERODINAMICI			TERRENO PIANEGGIANTE, SENZA VENTO			EFFETTI DELLE ALTURE	
				CARATTERISTICHE	DIAMETRO (cm)	PRESSIONE (kg/cm²)			COEFF. DI RESISTENZA	SUPERFICIE FRONTALE (cm²)	SUPERFICIE FRONTALE EFFETTIVA (cm²)	INDICE DI POTENZA	ANDATURA TURISTICA PER TUTTA LA GIORNATA	VELOCITÀ MASSIMA	VELOCITÀ COSTANTE IN SALITA	VELOCITÀ COSTANTE A RUOTA LIBERA
BICICLETTE STANDARD	BMX (BICICLETTA «FUORISTRADA» PER RAGAZZI)	13,6	55	COPERTONI CON SCOLPITURE MOLTO PRONUNCIATE	50,8	2,8	2,5 0,95	0,014	1,1	4552	5017	146	16,2	44,7	19,6	31,9
	BICICLETTA CONVENZIONALE EUROPEA CON CICLISTA IN POSIZIONE ERETTA	18,1	72		68,5	2,8	2,78 0,54	0,006	1,1	5109	5574	140	18,1	44,4	17,5	38,6
	BICICLETTA DA TURISMO (CICLISTA CON LE BRACCIA TESE)	18,1	72	A TALLONE	68,5	6,3	1,99 0,37	0,0045	1	3994	3995	100	21,1	50	19,6	44,6
	BICICLETTA DA CORSA (CICLISTA IN POSIZIONE COMPLETAMENTE CHINA SUL MANUBRIO)	9	72	TUBOLARI	68,5	7,4	1,57 0,24	0,003	0,88	3623	3159	77	23,6	54,5	20,9	50,2
MODELLI MIGLIORATI	COMPONENTI AERODINAMICHE (CICLISTA IN POSIZIONE COMPLETAMENTE CHINA SUL MANUBRIO)	9	72	TUBOLARI	68,5	7,4	1,48 0,24	0,003	0,83	3623	2973	73	24,1	55,6	20,9	51,8
	CARENATURA PARZIALE (ZZIPPER, CICLISTA IN POSIZIONE CHINA SUL MANUBRIO)	9,5	72	TUBOLARI	68,5	7,4	1,35 0,24	0,003	0,70	3809	2694	67	24,8	57,4	21	54,5
	RECUMBENT (EASY RACER)	12,2	72	A TALLONE	ANTERIORE 50,8 POSTERIORE 68,5	6,3	1,35 0,43	0,005	0,77	3530	2694	75	23,2	56,6	32,4	54,2
	TANDEM	19	72 CIASCUNO	A TALLONE	68,5	6,3	2,41 1,20 0,73 0,37	0,0045	1	4831	4831	66	24,5	58,9	20,9	56,6
	«TIRATA» (SEGUE DA VICINO UN'ALTRA BICICLETTA)	9	72	TUBOLARI	68,5	7,4	0,87 0,24	0,003	0,50	3623	1765	47	28,2	66	21,9	67,1
PRIMATISTI	BLUE BELL (DUE RUOTE, UN CICLISTA)	18,1	72	TUBOLARI	ANTERIORE 50,8 POSTERIORE 68,5	7,4	0,27 0,36	0,004	0,12	4645	557	27	36,2	94,3	20,8	124,6
	KYLE (DUE RUOTE, DUE CICLISTI)	23,6	72 CIASCUNO	TUBOLARI		7,4	0,65 0,33 0,5 0,25	0,003	0,2	6503	1301	24	37,5	0,1	22,5	112,5
	VECTOR SINGLE (TRE RUOTE)	30,8	72	TUBOLARI	ANTERIORI 50,8 POSTERIORE 68,5		0,23 0,46	0,0045	0,11	4236	464	29	35,1	98,5	18,2	145
	VECTOR TANDEM (TRE RUOTE)	34	72 CIASCUNO	TUBOLARI	61		0,28 0,41 0,80 0,40	0,0045	0,13	4366	557	23	41,2	116,7	20,9	174,4
LIMITI TEORICI	BICICLETTA PERFETTA (NESSUNA RESISTENZA AL ROTOLAMENTO, NESSUNA RESISTENZA OPPOSTA AL VEICOLO)						1,39 0	0	0,8	3530	2787	59	26,9	93	21,6	55,8
	CICLISTA SENZA RESISTENZA (LA RESISTENZA AL ROTOLAMENTO COMPRENDE IL PESO DEL CICLISTA)						0,60 0,37	0,0045	1,1	1115	1208	41	29,6	73,7	21,4	80,9
	RECUMBENT PERFETTA (RESISTENZA SOLTANTO AL CICLISTA)						0,33 0	0	0,6	1115	650	14	43,6	93,8	27	107,7
	BICICLETTA PERFETTA CON IL CICLISTA IN POSIZIONE PRONA (RESISTENZA OPPOSTA A UN CICLISTA PICCOLO MA ROBUSTO)						0,23 0	0	0,6	743	464	10	48,9	105	37,3	105
	PROFILO AERODINAMICO PERFETTO CON CICLISTA IN POSIZIONE PRONA						0,03 0	0	0,05	1301	65	1	93,8	202,6	41,2	280,8
	DIETRO A VEICOLO A MOTORE	19	72	DA MOTOCICLETTA PER CORSA SU STRADA		4,9	0 0,55	0,006			VARIA CON LA VELOCITÀ	23	47,3	473,1	20,3	?
	BICICLETTA LUNARE (TUTA SPAZIALE 6,8 kg)	11,3	72				0 0,06	0,0045			0	3	382,2	3822	126,2	?

squadra di quattro corridori può raggiungere una velocità superiore di circa 6,4 chilometri a quella di un ciclista isolato. Tipicamente, a parità di bravura, un gruppo di turisti in bicicletta può procedere a un'andatura da uno e mezzo a cinque chilometri più elevata di quella di un qualsiasi ciclista che proceda da solo. Quanto più il gruppo è numeroso (poniamo, fino a dodici elementi), tanto più veloce dovrebbe essere.

I venti artificiali creati dal traffico automobilistico di passaggio possono aumentare da uno e mezzo a cinque chilometri la velocità di un ciclista per periodi di quasi sette secondi. Quanto più grande è il veicolo che passa, tanto più notevole è l'effetto. Una corrente costante di traffico può permettere a un ciclista di mantenere una velocità da sei a 10 chilometri più elevata di quanto gli sarebbe altrimenti possibile a parità di energia.

Quando un ciclista pedala direttamente nella scia di un veicolo a motore, si possono raggiungere velocità veramente eccezionali. Il 25 agosto 1973 Allan V. Abbott, un medico californiano, stabilì un primato percorrendo un miglio alle Bonneville Salt Flats, nello Utah, alla media di 138,674 miglia (221,878 chilometri) all'ora. John Howard, un ciclista olimpico americano, sta cercando di battere il primato di Abbott e di raggiungere facendosi «tirare» da un veicolo a motore a una velocità superiore a 240 chilometri all'ora.

Sebbene i risultati che abbiamo descritto siano di per sé molto importanti, ci si chiede se avranno mai applicazione pratica al di là del loro effetto sui primati di velocità. Per gran parte dei ciclisti di tutto il mondo sembra improbabile che queste fatiche abbiano molta utilità immediata. Per esempio, nei molti paesi in via di sviluppo in cui la bicicletta è il principale mezzo di trasporto i ciclisti vanno a circa 11 chilometri all'ora, spesso con un carico pesante; la resistenza aerodinamica diventa più importante di altri impedimenti al moto delle biciclette soltanto a velocità superiori a 16 chilometri all'ora. Anche qui il lavoro sull'aerodinamica dà un contributo. Senza di esso i progettisti non saprebbero per quale ragione dovrebbero in gran parte ignorare l'aerodinamica per i veicoli a propulsione umana che si muovono piano.

Nel prospetto sono compendiate le prestazioni dei veicoli a propulsione umana. I valori elencati sotto le forze per ogni veicolo rappresentano rispettivamente la resistenza aerodinamica e la resistenza al rotolamento. Le ultime cinque colonne a destra indicano rispettivamente la potenza richiesta a 32 chilometri all'ora come percentuale del rendimento del cicloturista, la velocità in chilometri all'ora sull'arco di tutta la giornata con una produzione di 75 watt, la velocità massima per una produzione di 750 watt, la velocità costante in chilometri all'ora su pendenze fino al 5 per cento con una produzione di 300 watt e la velocità in discesa a ruota libera sulla stessa pendenza.

Il Vector Tandem è rappresentato qui in pianta e in proiezione verticale. È il «fratello» della Vector Single presentata sulla copertina di questo fascicolo. Con un input di poco più di 750 watt da parte di ognuno dei due ciclisti, il tandem stabilì nel 1980 il record di velocità sui 200 metri percorrendo la distanza alla media di 62,92 miglia (100,57 chilometri) all'ora, con partenza lanciata di oltre un miglio. Nello stesso anno il Vector Tandem percorse 40 miglia (64 chilometri) sulla strada interstatale della California alla media di 50,5 miglia (80,8 chilometri) all'ora.

Per le biciclette destinate a un procedere lento, ma sicuro, è logico ridurre la resistenza al rotolamento migliorando gli pneumatici e lastricando le strade. I progettisti dovrebbero anche ridurre il peso della bicicletta per facilitare l'andatura in salita. La recente introduzione negli Stati Uniti di «biciclette da montagna» è un passo verso la produzione di biciclette leggere abbastanza resistenti per strade accidentate o non lastricate.

La conoscenza acquisita con le recenti ricerche nel campo dell'aerodinamica dei veicoli mossi dall'uomo può essere direttamente utile in parecchi modi. Anche se per molti anni ancora continuerà probabilmente a essere la principale rappresentante della categoria per via del generale consenso, del basso costo, della semplicità e dell'affidabilità meccanica, la bicicletta standard offre molte opportunità di innovazione. Una carenatura anteriore semplice ed economica, per esempio, ne aumenterebbe notevolmente le prestazioni. Non è da escludere che la bicicletta *recumbent* venga usata in misura maggiore dai pendolari e dai turisti per via del suo rendimento e della sua comodità.

Un'ulteriore applicazione della tecnologia sarebbe quella di dotare una *recumbent* di un motore piccolo e leggero che servirebbe soprattutto per aiutare ad accelerare da fermo e a superare le salite. Dotato anche di tutta l'aerodinamicità compatibile con le esigenze di ventilazione e di stabilità, il veicolo sarebbe un vero ciclomotore. (I veicoli venduti attualmente sotto questo nome non sono dei veri ciclomotori, ma motociclette con motore di ridottissima potenza.)

La ricerca recente ha ispirato agli inventori parecchi veicoli a propulsione umana destinati a usi speciali. Paul Schöndorf, professore di ingegneria alla Fachhochschule di Colonia ha costruito, per le persone anziane e per gli handicappati, una serie di tricicli *recumbent* facilmente pedalabili e utilizzabili in qualsiasi condizione di tempo. Veicoli analoghi andrebbero bene anche per le comunità di pensionati. Douglas Schwandt, del Veterans Administration's Rehabilitation Engineering Research Development Center di Palo Alto, in California, ha costruito tricicli e biciclette per paraplegici mossi da manovelle. William Warner, un paraplegico che un tempo era stato detentore del primato per veicoli azionati con le mani nelle gare patrocinate dall'International Human Powered Vehicle Association, dice che una persona mutilata può spingere un veicolo del genere molto più velocemente di una normale sedia a rotelle e può quindi acquisire un nuovo senso di libertà e di mobilità. (L'attuale primato di 40,14 chilometri all'ora fu stabilito nel 1981 da Ascher Williams del centro di riabilitazione di Palo Alto.)

In linea di massima un veicolo a propulsione umana a profilo aerodinamico e completamente racchiuso potrebbe essere molto utile come mezzo di trasporto. Una persona potrebbe andare a velocità variabili da 30 a 50 chilometri all'ora in qualsiasi condizione di tempo. Così come sono progettati attualmente, però, questi veicoli non servirebbero sulle strade aperte. Essi mancano di sufficiente ventilazione, visibilità, manovrabilità e di certi elementi di sicurezza quali i fanali e i tergicristalli. Nella maggior parte dei casi, inoltre, non è facile salirvi o scenderne.

Produrre un veicolo di uso pratico di questo genere richiederebbe investimenti e sforzi ingegneristici paragonabili a quelli in gioco nella produzione di una nuova automobile. Anche così però il veicolo a pedali non sarebbe sicuro in un traffico che comprendesse un gran numero di veicoli a motore. È giocoforza concludere che un veicolo a propulsione umana interamente racchiuso non costituirà una forma pratica di trasporto fino a quando la penuria di carburante non toglierà dalle strade i veicoli motorizzati o fin quando non verranno costruite strade speciali per macchine a pedali.

Molto più probabile è lo sviluppo di automobili più leggere e più efficienti dal punto di vista del consumo di carburante che impieghino molta della tecnologia di cui si è parlato. Uno di noi (Malewicki) ha già costruito un veicolo del genere, una vettura monoposto del peso di 230 libbre (circa 104 chilogrammi), la quale detiene vari primati per quel che riguarda il risparmio di carburante, alla velocità massima di 55 miglia all'ora consentita sulle autostrade americane, con motore a benzina (157,2 miglia con un gallone, cioè 55,15 chilometri con un litro) e con motore diesel (156,3 miglia con un gallone, cioè 55 chilometri con un litro). Il primato diesel fu stabilito sul tratto Los Angeles-Las Vegas, percorso alla media di 56,3 miglia (90,08 chilometri) all'ora. Una tendenza verso tali vetture potrebbe rinviare il momento in cui al veicolo a propulsione umana verranno pienamente riconosciuti i meriti che gli spettano.

La biomeccanica per il record dell'ora

Come è noto, il regolamento della Federazione ciclistica internazionale impedisce che possano essere adottate, in gare ciclistiche, forme aerodinamiche particolari della bicicletta o applicate aggiunte aerodinamiche all'insieme «atleta-mezzo meccanico». Da questo punto di vista non è stato quindi possibile un sufficiente sviluppo di questo veicolo a propulsione umana.

Recentemente quando sono stato interessato, per quanto riguarda la parte biomeccanica, all'impresa di Francesco Moser di attaccare il record dell'ora, tenendo sempre presente i regolamenti internazionali, peraltro molto rigidi, ho cercato di utilizzare qualcosa che potesse sfruttare quanto di nuovo era stato recentemente realizzato nella tecnica.

A tal fine si è vista l'opportunità di realizzare un telaio con ruote di dimensioni diverse e precisamente con la ruota anteriore più bassa. In tal modo viene ridotta la resistenza aerodinamica della parte anteriore del veicolo. La ruota a raggio più piccolo, però, presenta una più elevata resistenza al rotolamento e, quindi, si è dovuto progettare un telaio capace di scaricare una maggiore percentuale del peso dell'atleta e della bicicletta sulla ruota posteriore, quella motrice, a diametro maggiore, così da ridurre il peso e l'impronta al suolo del tubolare della ruota anteriore. Inoltre, appunto per minimizzare il problema della resistenza al rotolamento, sono stati studiati e realizzati dalla Victoria, dietro nostre indicazioni, tubolari che presentano la minor resistenza possibile al rotolamento avendo una carcassa particolarmente rigida e sezione molto ridotta.

Per quanto riguarda la bicicletta, essa è stata realizzata in modo tale da avere la massima rigidità senza guardare eccessivamente alla riduzione del peso. Questo sembra in contrasto con quanto è stato precedentemente fatto in analoghi record. Sembra, infatti, che tutti coloro che si sono dedicati a tale argomento abbiano pensato alla riduzione del peso a scapito della rigidezza del veicolo mentre, al contrario, tutti i ricercatori e gli studiosi che in ambienti universitari hanno approfondito il problema del raggiungimento delle massime velocità a regime costante hanno sempre privilegiato l'aspetto aerodinamico piuttosto che l'alleggerimento estremo.

Prove aerodinamiche in galleria del vento di una delle biciclette approntate per il record dell'ora (Fotoservizi Franco Girella, Milano).

Ci è sembrato corretto questo modo di affrontare il problema in quanto, una volta raggiunta una determinata velocità, il mantenimento della stessa è assicurato, a parità di potenza muscolare erogata, più dalla riduzione della sezione frontale e dal miglioramento del coefficiente di penetrazione che dalla riduzione del peso.

La tecnologia moderna ha inoltre permesso l'utilizzazione di ruote costruite in materiali diversi da quelli tradizionali adottati fino a oggi, che sono stati, in successione di tempo, il legno, il ferro, l'acciaio, le leghe metalliche leggere. Oggi i materiali «compositi», ossia sostanze plastiche che possono essere facilmente plasmate nelle forme volute, risultano sufficientemente resistenti da permettere la realizzazione di ruote strutturate completamente con tali materiali. Questo in ossequio al regolamento internazionale che non consente l'applicazione, sulle biciclette, di aggiunte o di prolungamenti aerodinamici, ma che non proibisce l'utilizzazione di materiali nuovi. L'uso di materiali nuovi impone, talvolta, di adottare forme nuove. Ecco perché sono state proposte ruote a forma lenticolare in materiale composito e cioè ruote a superficie discoidale, senza raggi. È stata perciò la scelta stessa di questo nuovo tipo di materiale che ha imposto la realizzazione di ruote con una forma diversa dall'usuale. Infatti non potevano essere realizzati, con tale materiale, i tradizionali raggi, ma doveva essere necessariamente adottata una struttura con ruote a disco. Se poi il disco, come è naturale, soprattutto se è a profilo lenticolare, offre una riduzione della resistenza aerodinamica, in assenza di vento, questo costituisce un fatto «occasionale», gradito certamente, anche perché ottenuto in ossequio al rigidissimo regolamento.

Per quanto riguarda l'adozione di un casco che comprenda anche il collo oltre la testa, esso risponde a precise esigenze di sicurezza. Infatti, in caso di urto, l'uso di un casco che applichi e scarichi l'energia cinetica dell'urto su strutture di ampia superficie e di elevata resistenza, quali le spalle, la parte anteriore del torace, la parte posteriore del dorso, rappresenta un fattore di sicurezza per l'atleta. In tal modo, infatti, in caso di urto della testa contro la pista, l'effetto dell'urto non viene assorbito solamente dalle delicate strutture del collo, ma da una struttura ben più resistente quale il tronco. L'adozione di un casco così strutturato può essere accettata solamente in competizioni nelle quali l'atleta deve mantenere una posizione la più immobile possibile, senza rotazioni della testa sul collo, come nel caso di un record a velocità costante, dove non esista il problema di sorvegliare gli avversari. La soluzione di un tale tipo di casco, quindi, è valida solo per record come quello dell'ora o in altre situazioni analoghe.

Con questa serie di soluzioni si è voluto proporre qualcosa di nuovo in un ambiente in cui il veicolo di gara era rimasto fisso a una struttura standard fin dal secolo scorso. In tutti gli sport vi sono state innovazioni tecnologiche superiori a quelle adottate nel ciclismo. Il tentativo (coronato da successo il 19 e il 23 gennaio 1984) di Moser di attaccare il record dell'ora ha costituito l'occasione per verificare se era possibile compiere un passo in avanti. Questo passo in avanti è tutt'altro che sterile perché se, oggi, può vedere la sua pratica applicazione soprattutto nella massima espressione velocistica, non è detto che alcune delle soluzioni proposte non possano avere più larga applicazione o, comunque, costituire uno stimolo per uno sblocco di una situazione da troppo tempo cristallizzata.

Antonio Dal Monte

Mediante scavi rettangolari fu riportata in luce una delle tre fattorie individuate a Hascherkeller, una località della Bassa Baviera che risale all'Età del ferro. Il mutamento di colore che si osserva davanti al regolo di due metri che fornisce la scala rivela una parte del sistema di fossati perimetrali che serviva a separare questa fattoria da quella con essa confinante in direzione ovest. Gli agricoltori usavano erigere palizzate nei fossati per impedire agli animali domestici di uscire dalle aree recintate e, invece, agli animali selvatici di entrarvi di notte.

Il profilo di un fossato scavato a Hascherkeller tre millenni fa è caratterizzato da un'intrusione a *V* di humus scuro nel sottosuolo di loess giallastro, che ricopre la ghiaia di una terrazza fluviale. Il fossato era uno dei due fossati concentrici che circondavano la fattoria più a ovest e nei quali furono trovati frammenti di ossa, pezzi di intonaco di fango cotto, con cui erano stati ricoperti i muri dell'edificio, e cocci.

Una comunità agricola dell'Età del ferro in Europa

Scavi eseguiti in Baviera hanno portato in luce tracce di un'economia risalente al periodo compreso tra il 1000 e l'800 a.C. e caratterizzata dai primi scambi commerciali tra agricoltori e artigiani specializzati

di Peter S. Wells

La vita urbana era un fenomeno comune nel Vicino Oriente già attorno al 3500 a.C., ma per la maggior parte gli insediamenti europei a nord delle Alpi furono poco più che piccoli villaggi sino alla fine della tarda Età del bronzo e all'inizio della prima Età del ferro. In questa parte del mondo la transizione alla vita urbana cominciò ad aver luogo attorno all'800 a.C. Fu questo un periodo di rapido mutamento nell'Europa caratterizzata da un clima temperato, periodo che vide sia la crescita del commercio sia l'espansione della produzione di metalli. Le conoscenze archeologiche in proposito derivano principalmente dallo scavo di cimiteri, in numero di migliaia, e dalla scoperta casuale (ma non infrequente) di quantità di metalli sepolti. Solo alcuni insediamenti europei di questo periodo sono stati però studiati sistematicamente, cosicché gli sviluppi economici che condussero alla formazione delle prime città dell'Europa Centrale rimangono scarsamente noti. Qui descriverò i reperti trovati negli scavi, prolungatisi per quattro stagioni, nel sito di una fattoria di questo periodo, nella Bassa Baviera, e metterò quei reperti in relazione all'ascesa, in questa e altre località dell'Europa Centrale, di insediamenti di dimensioni maggiori: gli antecedenti delle città commerciali del periodo medioevale.

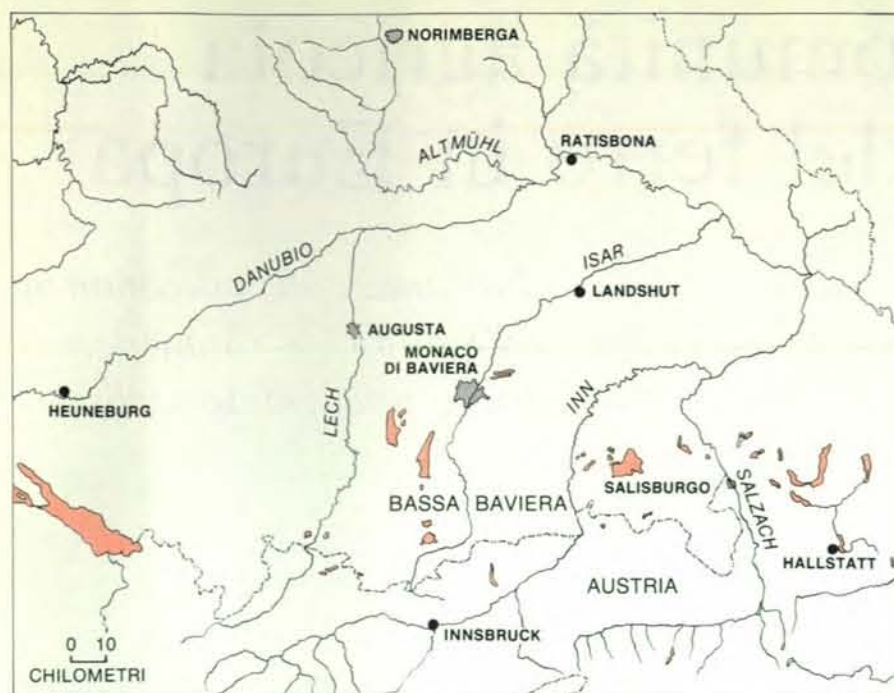
Il sito, Hascherkeller, si trova su una terrazza di sabbia e ghiaia che forma il confine settentrionale di una stretta valle fluviale alla periferia di Landshut, la principale città della Bassa Baviera. La terrazza si trova all'altezza di 15 metri al di sopra del fiume Isar, un affluente del Danubio, e il suo deposito di detriti glaciali trasportati qui dalle acque del fiume è ricoperto da uno spesso strato di loess, un suolo sedimentario di un colore giallo pallido attribuito a deposizione eolica alla fine dell'Epoca glaciale. Il loess è ricoperto a sua volta da un ricco humus, prodotto di millenni di trasformazione del suolo, e i 40 centimetri superiori dell'humus sono stati alterati dalla moderna aratura profonda. Nulla sopravvive perciò della superficie originaria del suolo dell'insediamento preistorico. Rimangono solo le tracce di scavi profondi compiuti dagli abitanti di questo sito nel periodo in studio, come buche di vario genere e fossati di confine: intrusioni scure di materiale humico nel loess altrimenti inalterato.

A cominciare dal 1978 i miei collaboratori e io iniziammo la divisione dell'area che intendevamo scavare in aree rettangolari, per lo più adiacenti, di cinque metri per dieci. Asportammo per mezzo di badili il terreno alterato dall'aratro nel corso di due scavi consecutivi, fino a una profondità di 20 centimetri in ciascuno scavo. Raccogliemmo e catalogammo i materiali archeologici rinvenuti nell'humus alterato secondo il livello e l'area di origine. Una volta raggiunto il fondo dell'humus smosso dall'aratro, continuammo il lavoro con zappe e cazzuole fino a portare in luce la superficie superiore del loess sottostante. Prima di procedere allo scavo dei materiali eterogenei in esso inclusi - che riempivano le buche e le fosse - eseguimmo un rilevamento cartografico e fotografico. L'humus contenuto in ciascuna buca fu poi suddiviso in due parti uguali e le due metà furono esaminate separatamente. Tutto l'humus fu estratto e setacciato con una rete metallica i cui fori erano di circa sei millimetri di diametro.

A mano a mano che il nostro lavoro proseguiva, diveniva evidente che l'insediamento era consistito in tre aree recintate l'una attigua all'altra ed estendentisi da est verso ovest; ciascuna risultava poi delimitata da un doppio fossato. Queste aree recintate erano ben definite sui loro confini settentrionale, orientale e occidentale, mentre l'erosione fluviale ne aveva distrutto il lato meridionale. Il nostro lavoro nel primo anno e nelle stagioni successive si concentrò sui contenuti dell'area occidentale e di quella di mezzo.

Benché l'aratura avesse distrutto la parte superiore dei doppi fossati, la parte che restava aveva una larghezza di ben tre metri. Gli scavi rivelarono che i fossati avevano una sezione a *V* ed erano profondi circa un metro e mezzo. Il materiale di colmata era bruno scuro e conteneva piccoli frammenti di vasellame e di ossa di animali. Una parte del fossato interno dell'area recintata occidentale presentava sul fondo una serie di 19 buche per pali, segno evidente che un tempo l'area era delimitata da una palizzata. La distanza media fra i centri di due buche successive era di 13,7 centimetri. Se tale distanza rappresenta il diametro medio dei singoli pali, si può presumere che la palizzata dovesse essere alquanto robusta.

Taluni esperimenti hanno indicato che pali di questo tipo piantati nel terreno marciscono entro pochi decenni. Finora non sono state ancora trovate altre buche per pali, ma molte parti dei fossati perimetrali presentano chiare tracce di nuovi scavi, necessari per mettere in opera altri pali quando quelli vecchi si deterioravano. Quest'osservazione suggerisce che tutti i fossati dovevano contenere palizzate. Lo scopo di queste opere doveva essere presumibilmente non tanto quello di proteggere gli agricoltori da eventuali aggressori, quanto quello di tenere gli animali domestici all'interno dell'insediamento e di impedire agli animali selvatici di entrarvi di notte.

La maggior parte delle informazioni che possediamo sulla comunità di Hascherkeller proviene dagli oggetti trovati nelle buche individuate all'interno delle aree recintate. Undici delle 21 buche più grandi da noi scavate poterono essere assegnate, sulla base dei loro contenuti, all'insediamento comprendente le tre aree recintate; le altre 10 appartenevano o a insediamenti più antichi e più piccoli dell'Età del bronzo o a posteriori occupazioni romane. Come i fossati, anche le buche, con il loro materiale di colmata

Il sito delle fattorie si trova a nord del fiume Isar, alla periferia della città di Landshut. In quest'area si trovano anche Hallstatt, un sito dell'Età del ferro, e Heuneburg, una città commerciale.

scuro, si distinguevano nettamente nei confronti del sottosuolo di loess. In effetti il loro materiale di colmata era ancora più scuro di quello dei fossati; ciò indicherebbe che contenevano una quantità maggiore di materia organica.

Le 11 buche si distinguevano per forma, dimensioni e contenuto in cinque categorie funzionali. Tre erano lunghe, strette e a forma di tazza ed erano orientate verso i punti cardinali della bussola: due in direzione nord-sud e una in direzione est-ovest. Si è trovato che tali orientamenti erano predominanti fra le case dello stesso periodo, scavate in altre località europee. Questa coincidenza suggerisce che le buche potessero assolvere la funzione di cantine per abitazioni: esse venivano usate probabilmente per riporvi i grandi vasi che proteggevano i generi alimentari dall'umidità, da fluttuazioni di temperatura e dalla predazione animale. I cocci trovati nelle tre buche corroborano questa conclusione: per la maggior parte si trattava infatti di frammenti di recipienti dalla parete molto spessa.

L'imboccatura di tre delle buche più profonde era circolare mentre le pareti erano quasi verticali. All'interno di buche simili, in altri siti dello stesso periodo, sono stati trovati semi di graminacee carbonizzati; inoltre esisteva un rivestimento con stuoie o con argilla, e ciò sta a indicare che erano state usate per immagazzinarvi cereali. Le tre buche di Hascherkeller non fornirono alcun indizio così preciso circa il loro uso; contenevano però alcuni cocci o altri oggetti di rifiuto e potrebbero avere assolto una funzione simile.

Altre due buche sarebbero state associate a un'attività metallurgica. Una conteneva una testa di martello in pietra e una forma di arenaria per la fusione di anelli. Lì vicino furono trovati molti piccoli frammenti di bronzo. Una buca adiacente conteneva ciottoli e terriccio colorati in rosso ed evidentemente stinti dall'azione di un calore intenso. Si potrebbe congetturare che, nella seconda buca, un fuoco molto caldo venisse usato per fondere il bronzo. La buca dove fu trovata la forma di fusione conteneva anche un peso per telaio di argilla refrattaria e cinque mulinelli per fuso in argilla. Questi reperti fanno pensare che in quest'area, oltre a fondere il bronzo, si esercitasse la tessitura.

Una grande buca, alcuni metri a nord dell'area recintata centrale, conteneva molto carbone di legna e sul suo fondo furono trovati i resti di una struttura in argilla simile a una cassetta. Più di metà di tutti i cocci trovati in questo sito furono tratti da questa buca. Per la maggior parte si trattava di cocci di oggetti che durante la cottura si erano rotti o deformati e che erano stati perciò gettati via. Pare inevitabile la conclusione che la struttura di argilla trovata nella buca fosse la camera di combustione di un forno di vasaio.

Le ultime due buche, poco profonde e con pareti dal pendio dolce, contenevano piccoli frammenti di vasellame e di ossa di animali. A quanto pare gli abitanti di Hascherkeller estraevano da queste buche loess che utilizzavano come materia prima per la produzione del vasellame o per rivestire di fango le pareti delle loro case. In seguito le buche furono gradualmente colmate con detriti. A questo proposito le 11 buche fornirono un totale di 198 chilogrammi di fango cotto, la cui cottura era avvenuta o in occasione di un incendio che aveva distrutto un edificio o quando il fango era stato usato per rivestire una parete adiacente a una qualche sorgente di calore, come un focolare. Il resto del fango era stato semplicemente ritrasformato in fanghiglia dagli agenti atmosferici dopo che gli edifici dell'insediamento erano caduti in disuso.

Le prove più abbondanti dell'insediamento umano a Hascherkeller sono quelle fornite dal vasellame rotto. Il numero totale di cocci recuperati fu di 14853. Nella grande maggioranza essi furono trovati nelle buche, ma 3828 furono rinvenuti nei fossati, negli strati di humus sovrapposti al sottosuolo di loess, e in altre aree. In ogni caso si trattava di oggetti semplici, rozzi, tipici della produzione ordinaria di terraglie dei contadini. Molte tombe di questo periodo nell'Europa centrale contengono belle ceramiche decorate, mentre meno del 4 per cento dei cocci di Hascherkeller presentano decorazioni di un genere qualsiasi.

Noi suddividemmo i cocci sulla base del loro spessore e trovammo che rientravano naturalmente in tre categorie. Alla prima appartenevano ceramiche dalla parete relativamente sottile, di meno di 4,5 millimetri di spessore. La seconda era formata da cocci di spessore compreso fra 4,5 e 9 millimetri, la terza di cocci di spessore superiore a 9 millimetri. Nella prima categoria rientravano i resti di piccole tazze, bicchieri e ciotole; essi erano anche quelli meno numerosi e quelli più spesso decorati. I cocci di ciotole più grandi e di giare alte, dalla grande imboccatura, con una finitura superficiale grossolana, costituivano la seconda categoria. I cocci della terza categoria appartenevano a giare dalla superficie rozzamente rifinita, usate soprattutto - come è presumibile - per conservarvi cibi. In altri siti di abitazione del periodo sono stati trovati molti di questi recipienti intatti, sepolti in buche e in cantine e contenenti ancora cereali.

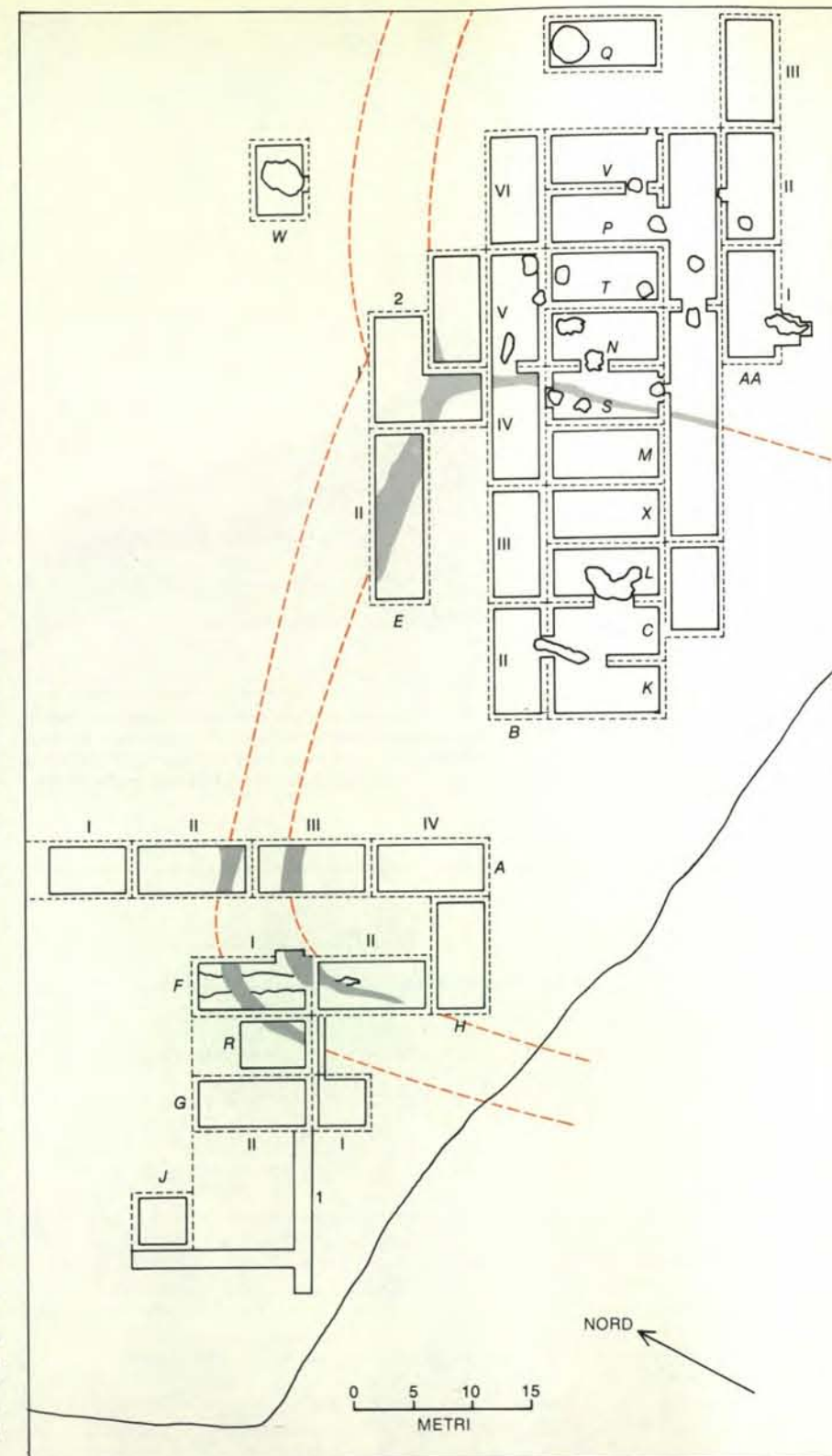
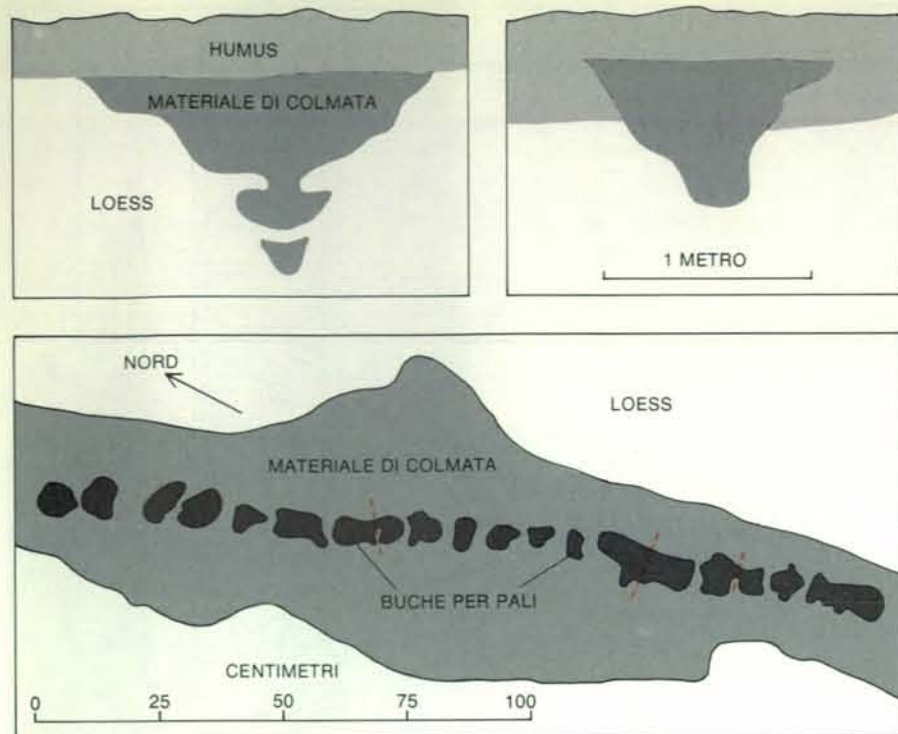
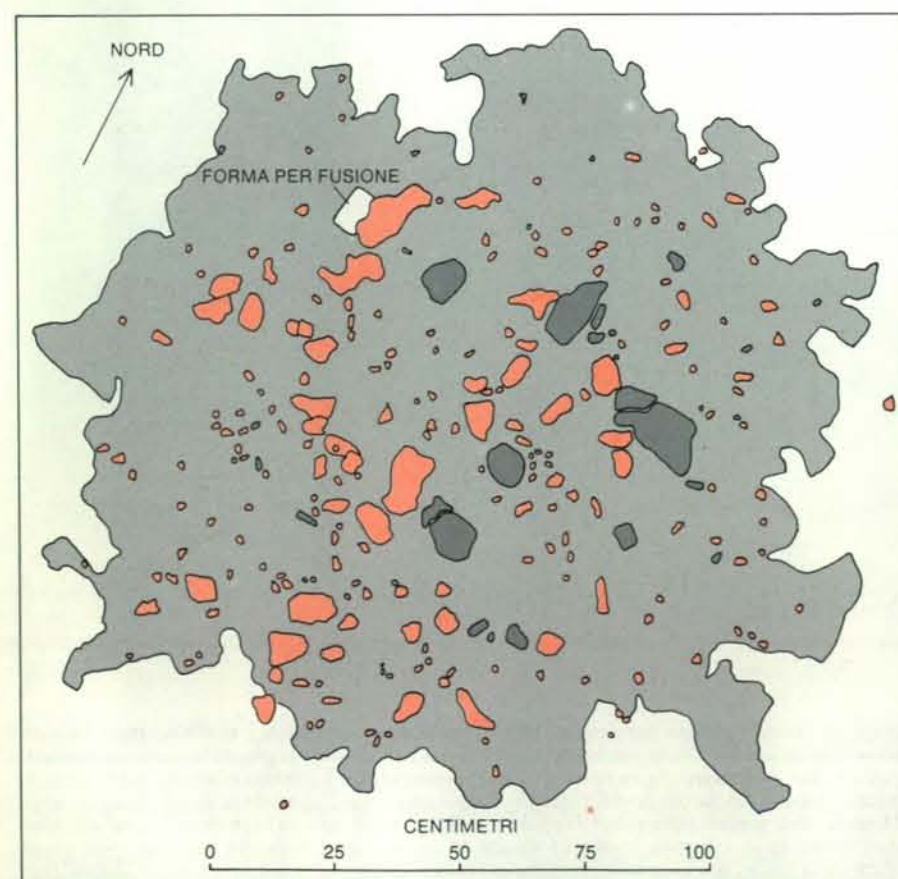
Se si fa eccezione per le grandi quantità di fango combusto, che fornirono informazioni utili sull'ubicazione di strutture edilizie nell'insediamento, i resti più numerosi a Hascherkeller furono frammenti di ossa di animali, ben 1435. Brenda Benefit, laureanda alla New York University, ebbe occasione di analizzarli e ha trovato che 253 di essi possono essere identificati come appartenenti a parti specifiche di animali di specie note. Le ossa identificabili sono in prevalenza (per l'87 per cento) di animali domestici; per il resto sono di animali selvatici. Fra le ossa di animali domestici predominano quelle dei maiali (37 per cento), seguite da quelle dei bovini (24 per cento) e da quelle degli ovini e dei caprini (33 per cento). Furono trovati inoltre frammenti di un piccolo numero di ossa di cavallo e di cane. L'animale selvatico più rappresentato fra i frammenti di ossa è il cervo rosso (*Cervus elaphus*), ma vi sono anche frammenti di ossa di lepre, porcospino e di una specie avicola (la quaglia). Inoltre c'è un buon numero di lische di pesce, ma nessuna che consenta l'identificazione della specie.

Brenda Benefit, dopo aver analizzato anche i denti di animali scoperti nel sito, sottolinea come l'usura dei denti di maiale si presenti in due modi diversi, il che starebbe a indicare che gli animali furono uccisi in due periodi distinti della loro vita. Molti erano stati macellati quando erano ancora porcellini da latte, non molto tempo dopo la loro nascita, gli altri press'a poco all'età di due anni. Queste sono le età alle quali vengono uccisi comunemente i maiali ancor oggi in Europa e che consentono di ottenere il massimo di carne in relazione alla quantità di cibo che si deve fornire agli animali quando, nei mesi invernali, non sono in grado di trovare da sé il proprio nutrimento.

Pecore, capre e bovini venivano lasciati vivere più a lungo, indipendentemente dal costo della loro alimentazione invernale. È presumibile che le pecore venissero allevate principalmente per la lana e le capre e le vacche per i latticini. I bovini potrebbero essere stati usati anche come animali da tiro e, quando infine venivano uccisi, come fonte di pelle e cuoio.

Oltre a questa varietà notevole di cibi carnei e di latticini, i resti di varie piante dimostrano che gli agricoltori di Hascherkeller consumavano anche cereali, verdure e cibi selvatici per arricchire la loro dieta. L'analisi dei resti di vegetali eseguita da Caroline Quillian Stubbs, laureanda alla Harvard University, indica che i principali cereali coltivati erano il miglio, il frumento e l'orzo; questi agricoltori coltivavano inoltre le lenticchie. Infine raccoglievano nocciole e parti di varie altre piante selvatiche che oggi sono considerate erbacce, ma che allora avevano una parte importante nella dieta dei primi europei. Gli abitanti di Hascherkeller raccoglievano l'attaccamani (*Galium*), il chenopodio (*Chenopodium*) e l'acetosa (*Rumex*).

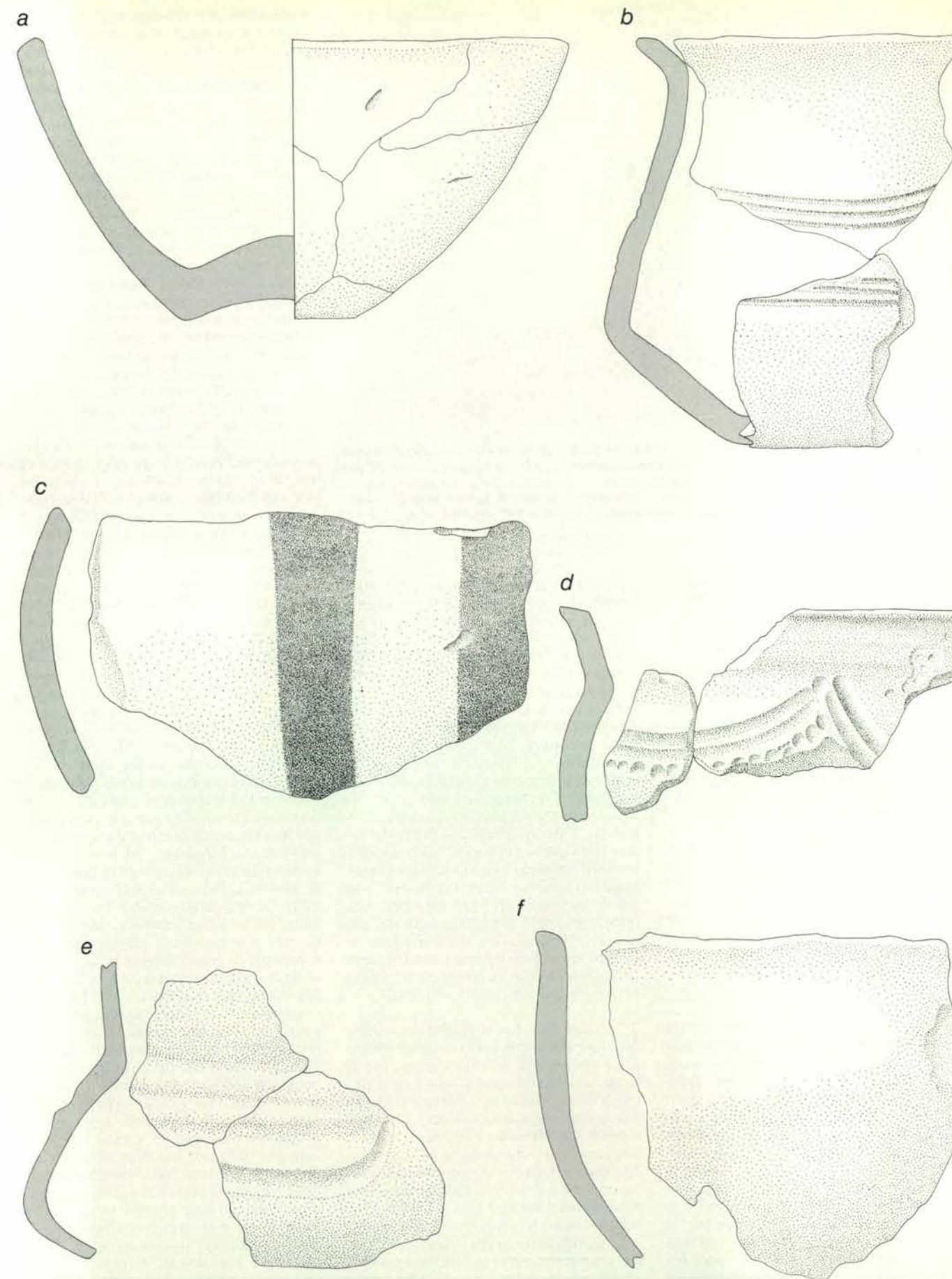
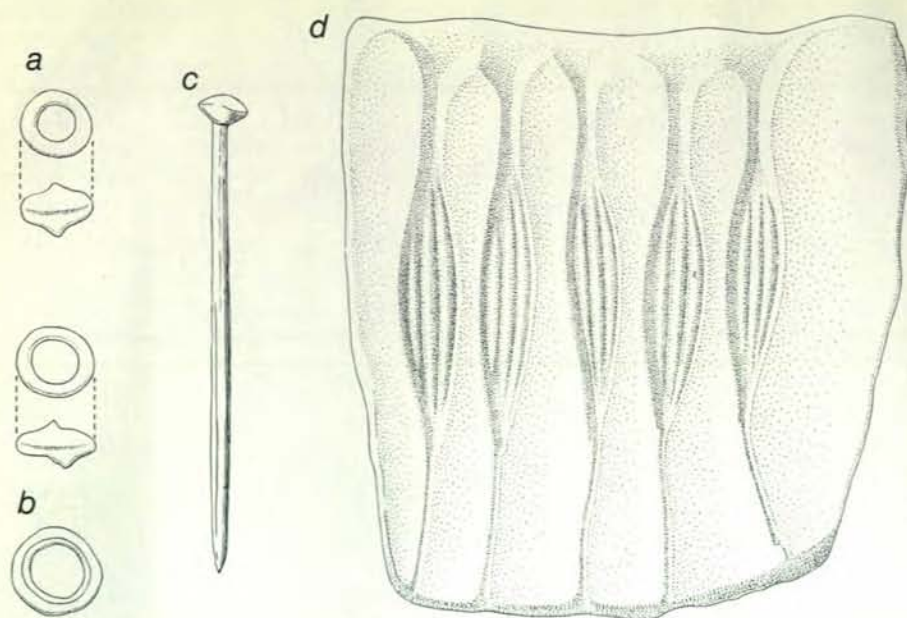
Quale immagine della vita nella prima Età del ferro si può ricavare da questi modesti reperti? Le dimensioni stimate delle tre aree racchiuse da fossati che formano le fattorie di Hascherkeller, ciascuna della superficie di circa 3000 metri quadrati, corrispondono alle dimensioni di singole fattorie individualmente recintate in altri insediamenti della tarda preistoria e dell'inizio dei tempi storici in Europa. È probabile che ogni fattoria fosse abitata da una famiglia formata da 5 a 10 individui e che comprendesse un'abitazione, un granaio per immagazzinarvi le provviste e strutture più piccole, come capannoni e laboratori. È evidente la contemporaneità delle tre fattorie: i loro fossati perimetrali coincidono esattamente e non si intersecano mai.

La vita quotidiana delle 15-30 persone, fra uomini, donne e bambini, che abitavano nell'insediamento può essere considerata sotto tre categorie economiche interconnesse: produzione di mezzi di sussistenza, produzione di oggetti di artigianato e commercio. Sotto la prima voce vi sono prove in abbondanza a conferma di un'organizzazione economica autosufficiente: allevamento di animali, che forniva carne e latticini, integrati da caccia e pesca, unitamente a produzione di cereali e di legumi, integrata a sua volta dalla

In questa pianta si vedono due delle tre fattorie, quella più occidentale e quella centrale. Le parti scavate dei doppi fossati che racchiudevano ciascuna fattoria sono in grigio; la parte non scavata, quale è stata evidenziata da un rilevamento magnetometrico, è indicata con linee tratteggiate in colore. Delle 21 buche più grandi (*linee di contorno continue*) riportate in luce dagli archeologi, 11 erano state scavate nella prima Età del ferro. Più di metà dei cocci venuti in luce nel sito sono stati trovati in una singola buca (*W*) situata all'esterno dei fossati che racchiudevano l'area recintata centrale; tale buca conteneva anche i resti di un forno per la cottura del vasellame. Due buche con funzione di cantine (*C* e *AA I*) contenevano invece le maggiori quantità di pezzi di fango cotto usato come intonaco nelle case, rispettivamente 75 e 44 chilogrammi. La buca più a occidente nel rettangolo *N* (*si veda l'illustrazione in basso nella pagina seguente*) conteneva una testa di martello in quarzite, una forma di arenaria per la fusione di anelli di bronzo, un peso in argilla per telaio, cinque mulinelli per fusi e 20 chilogrammi di frammenti di intonaco.

I fossati con sezione trasversale a forma di *V* (*a sinistra e a destra in alto*) si distinguono per la colorazione più scura dell'humus che li riempie rispetto al sottosuolo di loess, più chiaro. La vista in pianta di una parte del fossato interno dell'area recintata più occidentale (*qui sopra*) presenta 19 buche per pali su un tratto di 1,6 metri. Le linee tratteggiate in colore indicano buche doppie.

La buca «laboratorio» nel rettangolo *N* conteneva, oltre ad attrezzature per la fusione e per la tessitura, numerosi cocci (*in grigio scuro*) e frammenti di intonaco (*in colore*) in quantità da far pensare che avesse pareti intonacate con fango entro le quali ospitava artigiani tessili e per la lavorazione dei metalli. La forma di fusione in arenaria è indicata in alto a sinistra nella pianta.

raccolta di cibi vegetali selvatici. Che vi fosse o no un'eccedenza di carne o di cereali, è probabile che attività come la produzione di formaggi e la concia delle pelli procurassero all'insediamento merci che si conservavano facilmente in eccesso rispetto ai bisogni delle famiglie.

Quali merci, oltre a formaggio, cuoio e forse carne, potrebbero essere state prodotte in eccedenza? Di certo non il vasellame. A Hascherkeller ciascuna ciotola, tazza o giara di argilla veniva prodotta non con la ruota, ma con un procedimento laborioso, mediante argilla umida che veniva avvolta in rotoli a spirale e levigata poi con una spatola. È difficile che ciò potesse fornire a un insediamento agricolo un'eccedenza di prodotto. La nostra scoperta del peso per telaio e dei mulinelli per fuso, assieme alla prova che le pecore venivano allevate per anni prima di ucciderle, suggerisce però che una merce prodotta forse in eccesso rispetto al fabbisogno locale dovevano essere i tessuti in lana.

Si arriva così alla terza categoria economica: il commercio. Fra le prove che gli abitanti di Hascherkeller importavano materiali esotici c'è il fatto che artefatti di bronzo come gli anelli venivano prodotti localmente da rottami di bronzo fusi. I rottami venivano senza dubbio importati e anche varie spille di bronzo rinvenute potrebbero esserlo state. D'altra parte abbiamo trovato anche cinque oggetti di ferro frammentari, troppo piccoli e corrosi per poter essere identificati. Benché essi attestino che l'insediamento appartiene all'Età del ferro, non indicano necessariamente che quegli agricoltori importassero manufatti di ferro. Il minerale di ferro esiste quasi dappertutto, mentre non è così per le materie prime per la produzione della lega nota come bronzo, in particolare per lo stagno. Tracce della fusione del ferro in loco, sotto forma di una massa di scorie, confortano la conclusione che gli abitanti di Hascherkeller avessero familiarità con la lavorazione di questo metallo. Non ci sono, invece, indizi di una fusione locale del bronzo.

Fra le altre importazioni ci sono perline di vetro e grafite, quest'ultima usata per decorare la superficie di alcune terraglie di produzione locale. Se, sulla superficie di un recipiente, si sfrega grafite prima della cottura, si forma una vernice nera lucida. Questo tipo di ornamentazione, sotto forma o di un rivestimento completo della superficie o di una serie di bande, fu particolarmente diffuso nell'Europa centrale a cominciare dal 1000 a.C. circa. I principali depositi di grafite si trovavano ad almeno 100 chilometri da Hascherkeller, a est di Passau sul Danubio e verso nord in Boemia. Per quanto riguarda le perline di vetro, ne sono state rinvenute quattro, di colore verde-azzurro: due nella buca che conteneva il forno e due in un'altra buca. Non si sa dove siano state prodotte, ma nessun indizio a favore di una produzione locale di vetro è stato trovato né in quest'insediamento né in altri siti di questo periodo nell'Europa centrale. A giudicare dalla distribuzione

I cocci di sei recipienti sono inseriti nel profilo completo o parziale del recipiente originario. Quelli decorati hanno parete più sottile di uno dei due non decorati (*a*). Un recipiente (*c*) presentava bande nere, ottenute sfregando la superficie con grafite prima della cottura. Le miniere di grafite più vicine si trovavano a 100 chilometri di distanza ed è probabile che il vasaio locale si procurasse la grafite tramite scambi commerciali.

Fra gli oggetti di lusso scoperti a Hascherkeller c'erano perline di vetro (*a*), un anello di bronzo (*b*) e uno spillone pure di bronzo (*c*). Pur essendo stata trovata la forma per fusione (*d*) per produrre strisce con cui si ottenevano gli anelli, nessun anello prodotto in loco è venuto in luce. Lo spillone e l'anello di bronzo qui raffigurati, come pure le perline di vetro e la grafite usata per decorare il vasellame, furono ottenuti probabilmente barattandoli con eccedenze agricole.

relativamente vasta di tali manufatti, però, pare probabile che più di un centro di produzione di perle di vetro esportasse regolarmente le merci prodotte nelle comunità agricole della regione.

Ciò che rende quest'importazione di rottami di bronzo, di grafite e di perline di vetro particolarmente significativa in relazione al posteriore sviluppo delle città dell'Età del ferro nell'Europa a nord delle Alpi è il fatto che tutt'e tre i beni importati erano prodotti di lusso. La produzione di falci metalliche o di altri attrezzi agricoli a Hascherkeller, anche se in essi sono forse troppo accentuati certi tratti di eleganza, aveva uno scopo utilitario. Nessun contadino ha, invece, realmente bisogno di anelli e spille di bronzo, di vasellame decorato con la grafite o di perline di vetro verdi-azzurre per poter allevare più maiali o seminare più miglio. L'insediamento commerciava con il mondo esterno non per necessità ma per procurarsi ornamenti.

In quale periodo si situa quest'occupazione, e per quanto tempo durò? Sei campioni di carbone di legna tratti dalle tre buche fornirono, con il metodo del carbonio 14, date indicanti che il sito fu occupato per un periodo di 200 anni, fra il 1000 e l'800 a.C. Il vasellame e gli oggetti in bronzo dell'insediamento assieme alle perline di vetro importate corrispondono a una fase archeologica dell'Europa centrale nota come Hallstatt *B*. Quando si datano materiali appartenenti alla fase di Hallstatt *B* attraverso il confronto con materiali di culture mediterranee per le quali esistono documenti storici, tale fase viene collocata nei tre secoli compresi fra il 1000 e il 700 a.C. Tanto la cronologia assoluta quanto quella relativa sono dunque in stretto accordo.

Per quanto concerne la durata dell'insediamento, le prove sono meno dirette. I fossati furono rinnovati, e ho suggerito che questi lavori si resero necessari per ricostruire palizzate andate in rovina. Poiché i pali delle palizzate devono aver resistito alcuni decenni, l'insediamento non può essere durato di meno. Un'altra indicazione della durata dell'occupazione proviene dall'intonaco di fango preservato dalla cottura. Molti pezzi di intonaco dimostrano che i muri di alcuni edifici erano stati reintonacati e ridipinti due o tre volte. Ciò suggerisce che il sito sia stato occupato per almeno due o tre generazioni. L'occupazione non potrebbe essere stata anche più lunga? Sia la datazione con il carbonio 14 sia l'analisi del vasellame indicano un limite massimo di non più di due secoli. Il vasellame presenta, infatti, uno stile completamente omogeneo. È difficile credere che l'insediamento avrebbe potuto sottrarsi alle tendenze di mutamento che, in genere, si osservano in periodi che superino i 200 anni.

L'immagine di Hascherkeller che emerse dopo gli scavi è quella di una comunità economicamente autosufficiente, ma al tempo stesso connessa al mondo più ampio dell'Europa centrale per i piccoli lussi. Per soddisfare questi desideri c'era una sola via: il commercio, che è però una via a due sensi. Quel che gli agricoltori potevano offrire ai loro partner commerciali erano i prodotti delle loro fattorie: tessuti o filati di lana, formaggio, forse burro, carne salata e cuoio (o prodotti di pelle finiti).

In questo periodo era in corso nell'Europa centrale un'intensificazione generale della produzione agricola. Che l'insediamento di Hascherkeller sia rimasto o no attivo per due secoli interi, altri insediamenti furono occupati nello stesso periodo altrettanto a lungo e anche di più. La concimazione, la coltivazione a maggese e la rotazione delle colture conservavano la fertilità del suolo. Oltre alle falci venivano prodotti altri utensili di metallo, come asce, seghe, scalpelli e martelli.

Immaginiamo per un istante che l'espandersi della metallurgia avesse condotto allo sviluppo di associazioni di artigiani che non svolgevano attività agricole. Come rivelano le tombe, non si producevano però solo utensili in metallo. Accanto ai morti sono stati trovati armi e oggetti domestici decorati: spade, elmi, recipienti di bronzo e ornamenti d'oro. Come si procuravano di che vivere coloro che li producevano? Forse scambiando i loro prodotti con le eccedenze agricole delle fattorie, la cui produttività era in continuo aumento. Si può immaginare addirittura che esistessero intermediari che facilitassero con i loro viaggi gli scambi commerciali.

I reperti archeologici forniscono un esempio concreto di una tale associazione, sorta a non più di 160 chilometri da Hascherkeller. Presso le miniere di sale di Hallstatt, una comunità con una popolazione di circa 200 persone, fra l'800 e il 400 a.C., dedicò le sue energie esclusivamente all'estrazione e al commercio del sale. La quantità eccezionalmente grande di merci importate che accompagnava i morti di Hallstatt nella tomba è una prova eloquente del successo dell'esperimento di produrre e commerciare in comune una sola merce. I primi minatori cominciarono a lavorare alle miniere di sale di Hallstatt attorno al 1000 a.C., ma potrebbero essere stati agricoltori che estraevano sale per proprio uso, così come, nei secoli seguenti, i fonditori di bronzo, i tessitori, i pecorai e produttori di formaggi di Hascherkeller svolsero tali attività per proprio uso.

In ogni caso la comparsa della città mineraria di Hallstatt non fu un fenomeno unico. Piccole città con una popolazione di centinaia di persone dedite alla fusione e alla forgiatura del ferro sorsero nella regione alpina di confine dell'attuale Slovenia. Più vicino a Hascherkeller, nella Germania sudoccidentale, Heuneburg, un sito ben studiato, divenne un centro commerciale: un agglomerato compatto di grandi edifici in legno non diverso dalle prime città commerciali del Medioevo che sarebbero sorte 14 secoli dopo. Gli artigiani di Heuneburg si impegnarono in vari tipi di produzione primaria anziché in una qualche specialità singola, e lo stesso vale per città commerciali analoghe dell'Europa centrale. L'attività principale di queste città era però il commercio. Nessuno di questi agglomerati urbani avrebbe potuto sorgere senza il sostegno delle migliaia di piccoli insediamenti agricoli come Hascherkeller, dotati della capacità e della volontà di produrre eccedenze agricole sempre maggiori da scambiare con le città. Nella *Tempesta* di Shakespeare, Antonio dice: «Ciò che è passato è il prologo». A Hascherkeller vediamo il prologo della prima Età del ferro all'urbanesimo del Medioevo e del Rinascimento, che finirono con il plasmare il nostro mondo moderno.

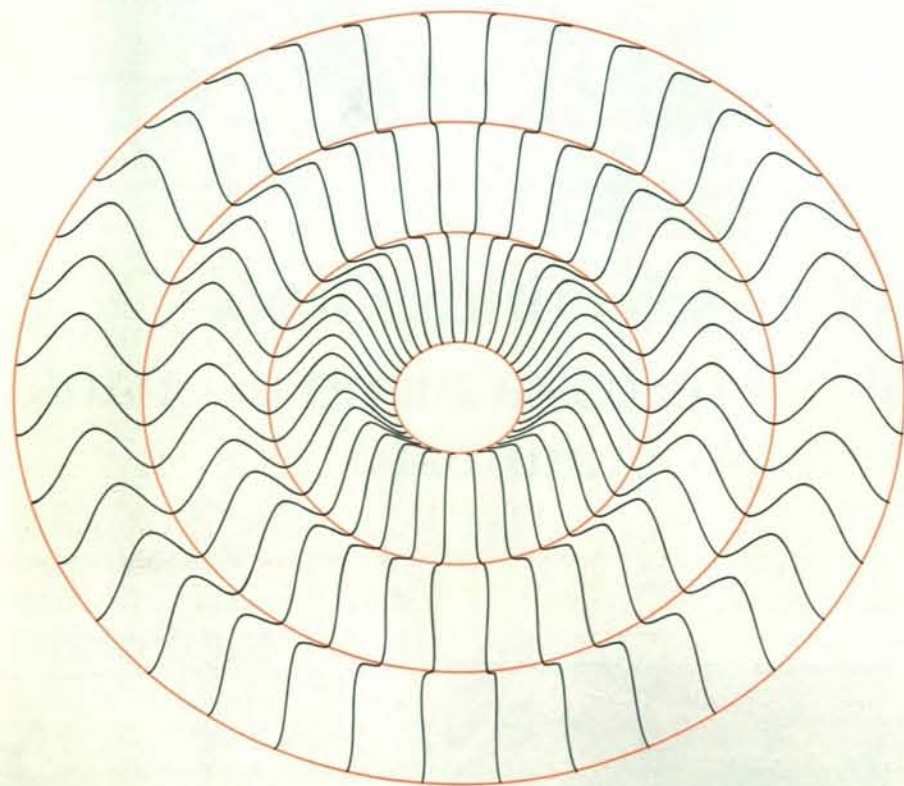
L'interpretazione delle illusioni visive

Evidentemente il sistema visivo organizza le immagini retiniche ambigue sulla base di regole di inferenza che sfruttano talune regolarità presenti nel mondo esterno

di Donald D. Hoffman

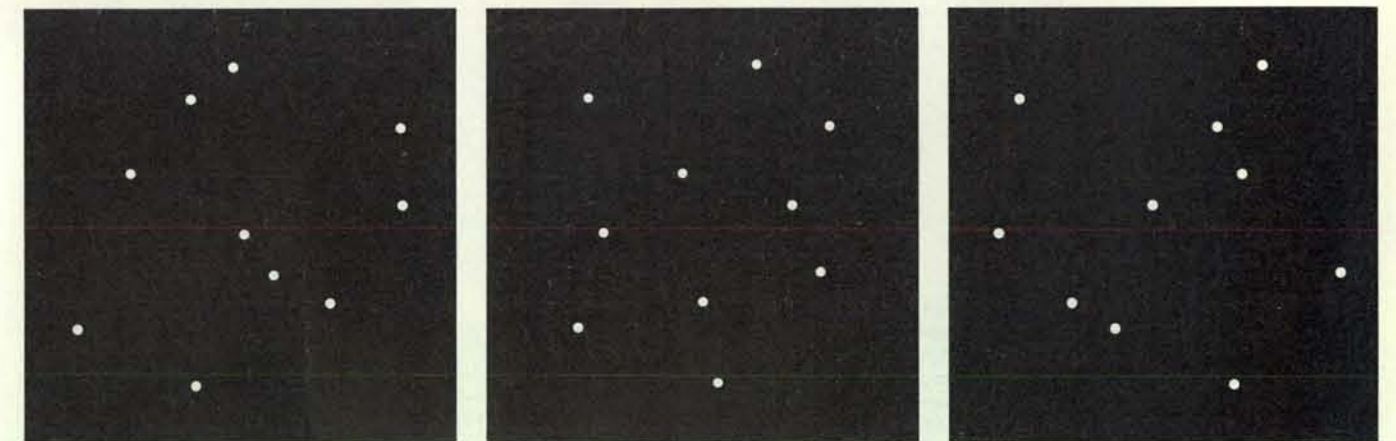
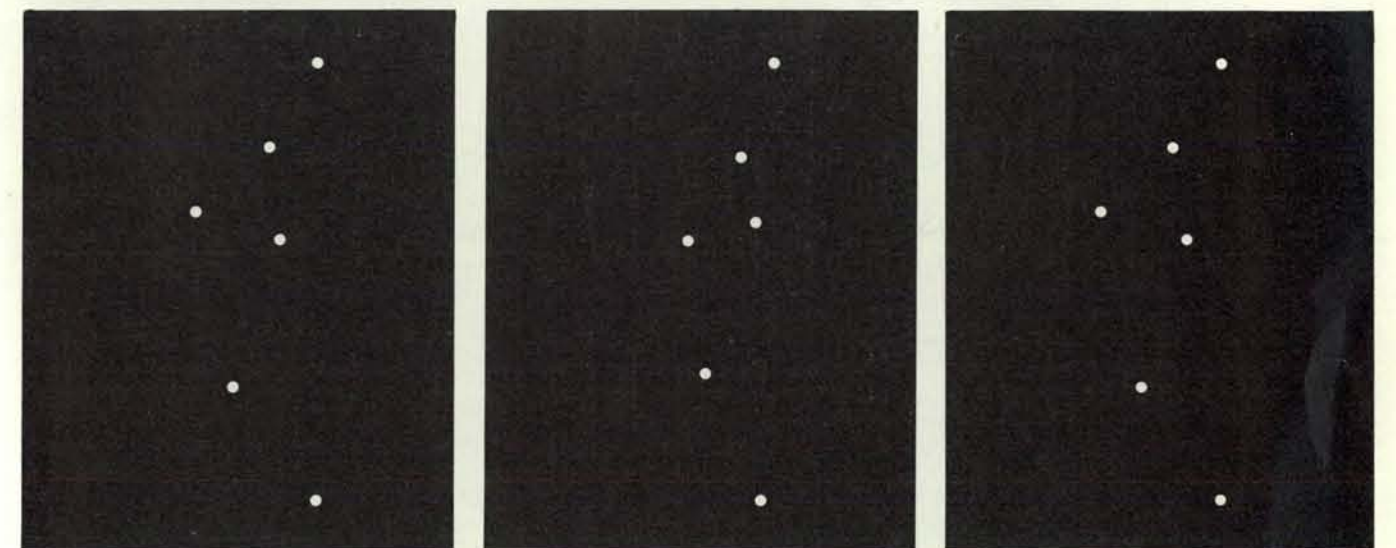
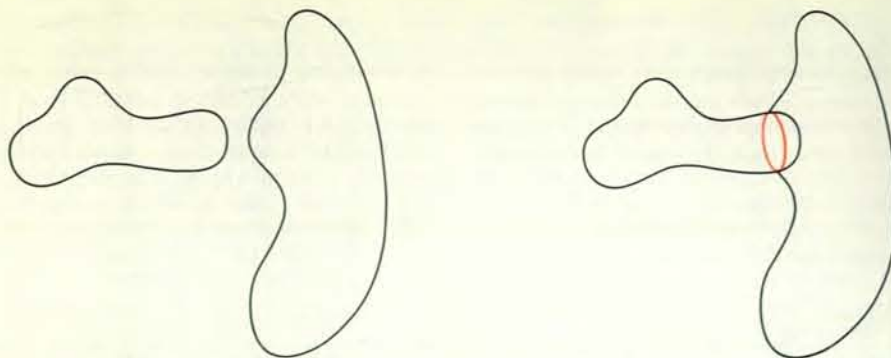
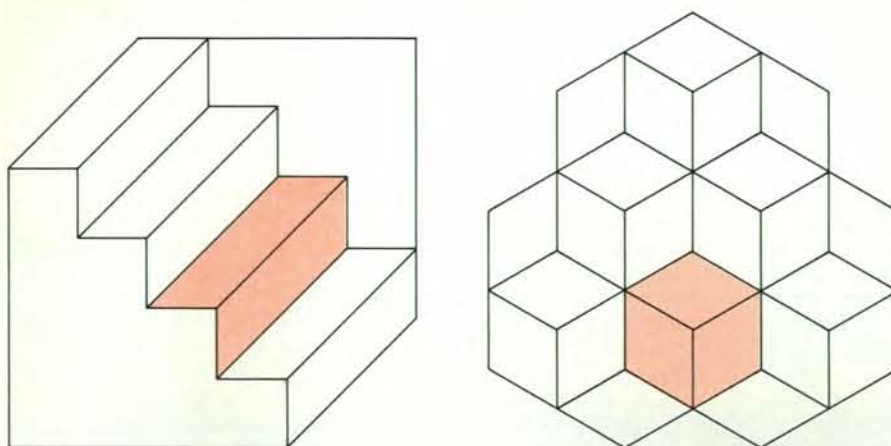
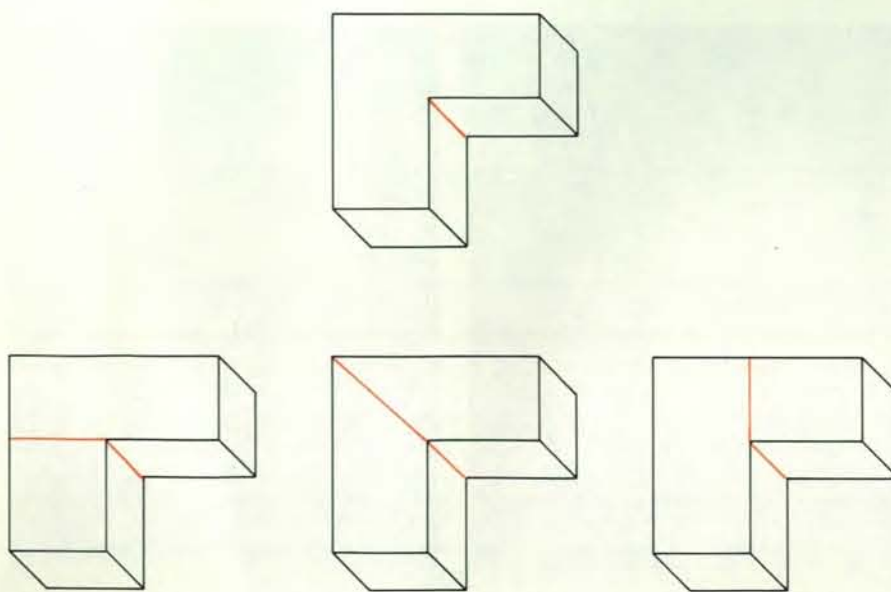
La visione è un processo di inferenza. Quello che vediamo quando guardiamo intorno a noi non dipende solamente da quello che c'è da vedere, ma anche dal modo in cui il nostro sistema visivo organizza e interpreta le immagini che si formano sulla retina dei nostri occhi. Una dimostrazione (che lascia molto sorpresi) di questo aspetto della percezione ci è fornita dalla superficie apparente, che si forma facendo ruotare il disegno di un'onda cosinusoidale attorno a un asse verticale, vista obliquamente (*si veda l'illustrazione in questa pagina*). Alla prima occhiata, l'immagine appare organizzata in un insieme di anelli concentrici in rilievo, con i margini fra gli anelli delineati approssimativamente dalle isolinee circolari in colore. Se però si rovescia la pagina, l'organizzazione cambia: ora le isolinee in colore, invece di trovarsi nei ventri fra due anelli, sembrano evidenziarne le creste. (Provare, per credere.) Chiaramente il sistema visivo fa qualcosa di più che trasmettere passivamente segnali al cervello: prende parte attivamente nell'organizzarli e nell'interpretarli.

Questa prima scoperta fa sorgere tre domande. In primo luogo, perché il sistema visivo deve organizzare e interpretare le immagini che si formano sulla retina? In secondo luogo, perché in questo processo resta fedele al mondo reale? E infine, quali regole di inferenza segue in questo processo? Per rispondere a queste domande è necessario un esame più approfondito delle figure di questo tipo.

Uno dei motivi per cui il sistema visivo deve organizzare e interpretare le immagini retiniche è semplicemente dovuto al fatto che vi sono molte possibili configurazioni diverse, nel mondo reale, coerenti con qualunque immagine retinica data. In altre parole, le immagini retiniche debbono essere organizzate e interpretate perché sono fondamentalmente ambigue. La loro ambiguità è dovuta in parte al fatto che il mondo è in tre dimensioni, mentre le immagini sulla retina sono essenzialmente bidimensionali. Per descrivere il mondo in tutta la sua ricchezza tridimensionale debbono necessariamente entrare in gioco alcune inferenze molto raffinate, da parte del sistema visivo, inferenze che per la più parte vengono tratte senza che noi ne abbiamo coscienza. Per esempio, la superficie cosinusoidale qui a fianco, come l'immagine retinica che ne abbiamo, è in due dimensioni. Tuttavia ci appare, e l'impressione è davvero convincente, come tridimensionale. L'impressione di profondità è completamente frutto di un'inferenza, o, per dirla in un altro modo, è un'allucinazione. Se, come io credo, allucinazioni di tal fatta non sono l'eccezione ma la regola, e se in effetti sono un aspetto necessariamente concomitante della percezione visiva, come si può giustificare la nostra fiducia nella percezione? Come si può ancora sostenere che in generale vedere significhi credere?

Si ottiene una superficie ambigua facendo ruotare un'onda cosinusoidale attorno a un asse verticale. La superficie inizialmente appare organizzata in anelli concentrici in rilievo, con le isolinee circolari in colore giacenti nei ventri delle onde fra gli anelli. Quando si rovescia la pagina, però, l'organizzazione appare mutata: le isolinee in colore sembrano tracciare le creste degli anelli.

Per capire i processi della visione, quindi, è necessario spiegare perché queste inferenze visive di solito intrattengano un rapporto non arbitrario con il mondo reale. Ci si apre una direzione di ricerca promettente, quando osserviamo che il mondo visibile è ben lungi dall'essere completamente caotico, e obbedisce invece a determinate leggi e presenta numerose regolarità. Se il sistema visivo è adattato a sfruttare queste leggi e queste regolarità nell'organizzazione e nell'interpretazione delle immagini retiniche, e se in qualche modo è vincolato a dare la preferenza all'interpretazione più credibile, data l'immagine e una conoscenza di tali leggi e regolarità, allora sarebbe possibile capire come mai le nostre allucinazioni visive intrattengano un rapporto non arbitrario, e addirittura utile, con il mondo esterno.

Un esempio particolarmente chiaro di questa impostazione ci viene offerto dalle ricerche sulla percezione visiva del movimento, effettuate da Shimon Ullman del Massachusetts Institute of Technology. Ullman ha esplorato la notevole capacità del sistema visivo umano di percepire la corretta struttura tridimensionale e il movimento di un oggetto a partire esclusivamente dalla sua proiezione bidimensionale in movimento, una capacità che Hans Wallach e Donald N. O'Connell dello Swarthmore College chiamano «effetto di profondità cinetica». Per esempio, se si fa ruotare in una stanza completamente al buio un pallone da spiaggia trasparente, sulla cui superficie siano state disposte in modo casuale delle lampadine, si percepisce immediatamente la corretta disposizione sferica delle luci (*si veda l'illustrazione superiore in questa pagina*). Quando il pallone smette di ruotare, viene a mancare anche la percezione della disposizione sferica. Come mai si vede la giusta struttura tridimensionale, quando la proiezione retinica bidimensionale in movimento è coerente con un numero infinito di possibili strutture tridimensionali? Ullman ha mostrato matematicamente che, se il sistema visivo sfrutta le leggi della proiezione, e se sfrutta il fatto che il mondo contiene oggetti rigidi, allora in linea di principio si può ottenere un'interpretazione unica e corretta. In particolare ha mostrato che tre viste di quattro lampadine non complanari sono sufficienti per risolvere il problema. Il punto chiave è che una regola di inferenza, basata su una legge (la legge della proiezione) e una regolarità (il fatto che il mondo comprende oggetti rigidi), consente al sistema visivo di formulare un'interpretazione corretta.

A questo stadio, tuttavia, ci si presenta un rompicapo. La stessa precisione matematica che mostra come la regolarità della rigidità sia sufficiente in linea di principio per interpretare la palla rotante

Quando le tre figure a puntini rappresentate qui vengono mostrate in rapida successione, si vede una sfera in rotazione. Il sistema visivo sembra adottare, per i puntini in movimento, l'interpretazione tridimensionale più rigida coerente con le proiezioni bidimensionali.

Quando queste figure di punti vengono mostrate in rapida successione, si vede una persona che cammina. In questo caso il sistema visivo sembra adottare l'interpretazione tridimensionale più rigida e planare coerente con i movimenti bidimensionali dei punti. La visualizzazione è basata su un esperimento condotto da Gunnar Johansson dell'Università di Uppsala, in cui erano state fissate piccole lampadine alle giunture principali (spalla, gomito, polso, anca, ginocchio e caviglia) di una persona, ripresa poi mentre si muoveva in una stanza al buio.

La trasversalità, un tipo di regolarità che si osserva comunemente nel mondo esterno, è alla base di una spiegazione unica per diverse illusioni visive. In base alla regola della trasversalità (definita da Whitman A. Richards e dall'autore), quando due superfici penetrano l'una nell'altra in modo casuale, si incontrano sempre in corrispondenza di una discontinuità concava (*qui in colore*).

La regola di partizione basata sulla regolarità della trasversalità è dimostrata con l'aiuto di queste due figure che subiscono una inversione quando vengono osservate attentamente abbastanza a lungo. In ambedue i casi i confini apparenti delle diverse parti della forma percepita si modificano quando la «figura» diventa lo «sfondo» e viceversa. Per esempio, nel caso dell'illusione della scala reversibile (*a sinistra*), pubblicata per la prima volta da H. Schröder nel 1858, le due strisce in colore, che secondo una interpretazione appaiono parti del medesimo scalino, improvvisamente sembrano parti di due scalini successivi, quando la scala subisce l'inversione. Analogamente, nell'illusione della pila di cubi (*a destra*) le parti romboidali in colore possono essere interpretate come facce di un unico cubo o, quando si verifica l'inversione, come facce di tre cubi distinti.

Questi blocchi a forma di *L* mostrano che la regola di partizione delle forme in corrispondenza di discontinuità concave è opportunamente prudente. La regola non definisce un contorno chiuso sul blocco in alto perché sembrano possibili tre diverse suddivisioni percepite, come illustrato in basso.

mostra anche come la regolarità della rigidità sia di per se stessa insufficiente per interpretare una analoga presentazione visiva. Questa presentazione è stata escogitata da Gunnar Johansson dell'Università di Uppsala come esempio di quello che egli chiama moto biologico (si veda l'articolo *La percezione visiva del movimento* di Gunnar Johansson in «Le Scienze», n. 86, ottobre 1975). Johansson ha attaccato piccole lampadine alle giunture principali di un uomo e ha registrato con una cinepresa i movimenti dell'uomo in una camera completamente al buio. Un singolo fotogramma del filmato ha l'aspetto di una collezione casuale di punti bianchi su uno sfondo nero. Quando il filmato viene proiettato normalmente, però, si vede immediatamente la corretta struttura tridimensionale dei puntini e si riconosce che c'è una persona invisibile che sta camminando (*si veda l'illustrazione in basso della pagina precedente*).

Quando il mio collega Bruce E. Flinchbaugh, ora ai Bell Laboratories, e io abbiamo preso in considerazione questo problema, ciò che ci lasciava perplessi era la possibilità di vedere la corretta struttura tridimensionale anche se, secondo i risultati di Ullman, ci mancano le opportune informazioni per una simile identificazione. Per inferire una struttura tridimensionale corretta sulla base della regolarità della rigidità è necessario avere almeno tre istantanee di quattro punti non complanari in una configurazione rigida. Nelle presentazioni del moto biologico, invece, nel migliore dei casi solo coppie di punti sono collegate rigidamente: per esempio la caviglia e il ginocchio o il ginocchio e l'anca. Le quadruple rigide di punti semplicemente non esistono.

La regolarità della rigidità, dunque, di per sé è insufficiente e ci porta a chiederci: quale ulteriore regolarità potrebbe essere sfruttata dal sistema visivo? Dopo aver battuto varie piste false, ci è venuta in mente una regolarità anatomica che potrebbe funzionare bene allo scopo. Per lo più, negli animali gli arti che debbono sostenere il peso sono vincolati, per la costruzione stessa delle loro articolazioni, a oscillare, nell'andatura normale, in un piano. Abbiamo dato, a questo vincolo, il nome di regolarità della planarità.

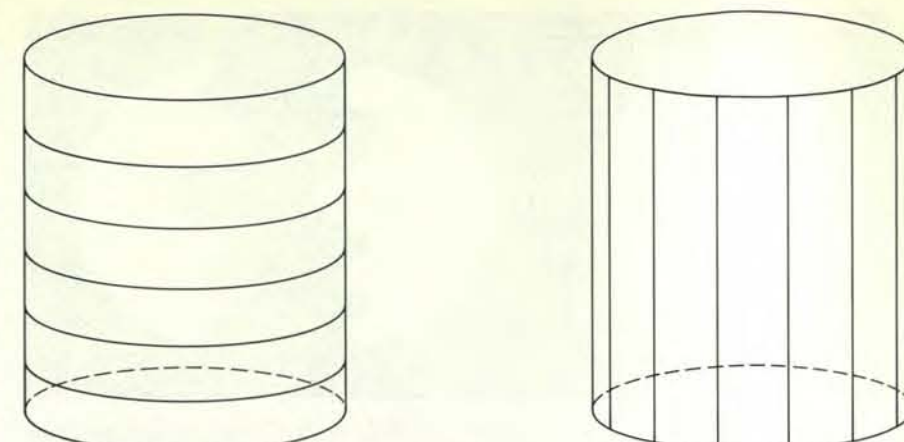
In effetti, la regolarità della planarità è sufficiente per interpretare correttamente le presentazioni del moto biologico. La struttura tridimensionale corretta può essere inferita o da tre istantanee di due punti che oscillano rigidamente in un piano o da due istantanee di tre punti (come una caviglia, un ginocchio e un'anca) che formano coppie rigide e oscillano in un piano. Questi risultati si accordano esattamente con l'osservazione di Johansson, secondo la quale bastano due o tre fotogrammi dei suoi filmati perché i soggetti percepiscano correttamente il moto biologico. Inoltre si ha non solo che tutti i moti tridimensionali governati dalla regolarità della planarità possono avere una interpretazione corretta, ma anche che, ogniqualvolta si trova una interpretazione per il moto delle immagini basata sulla regolarità della planarità o della rigidità, l'interpretazione stessa è corretta.

In breve, la probabilità che l'interpretazione sia errata è nulla, ipotizzando che la risoluzione dell'immagine sia infinita, o di poco maggiore di zero per una risoluzione meno che perfetta. Pertanto le strutture non rigide non possono travestirsi da strutture rigide, e i movimenti non planari non possono essere visti erroneamente come planari. Ancora una volta leggi e regolarità dimostrano il loro ruolo centrale nella spiegazione del modo in cui il sistema visivo raggiunge un'interpretazione unica e corretta di un'immagine retinica.

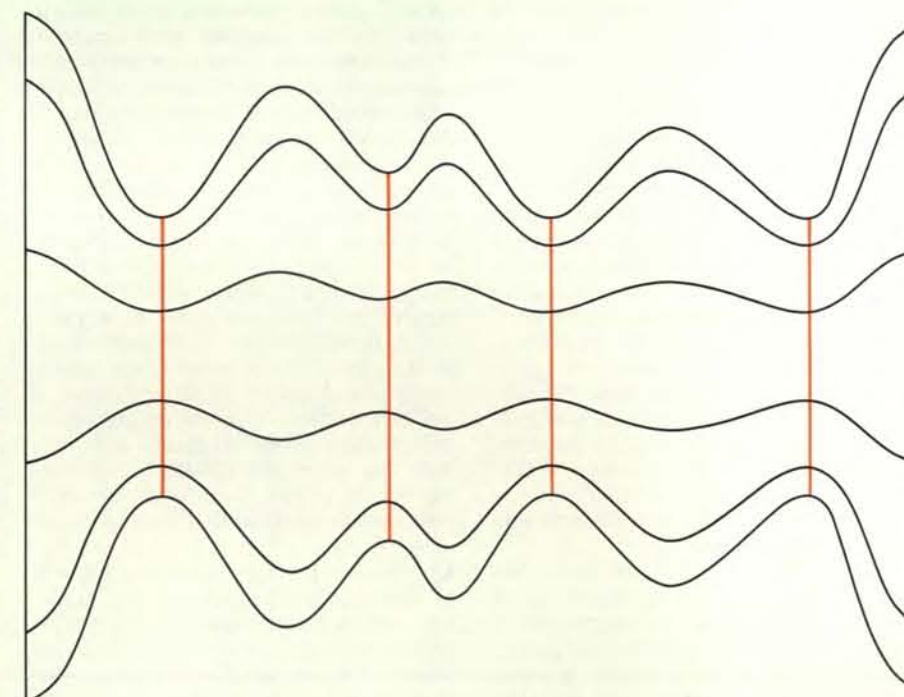
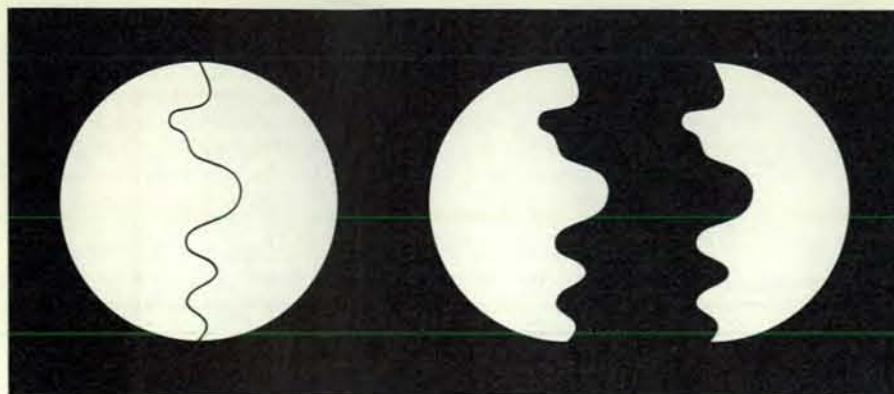
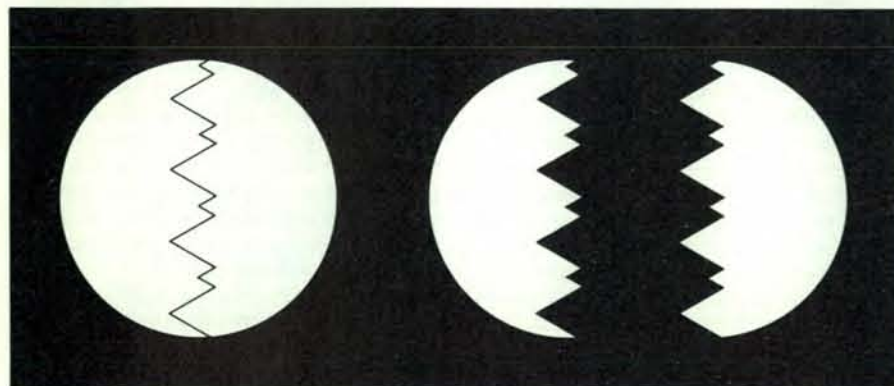
Torniamo ora alla superficie cosinusoidale. Il suo principale motivo d'interesse sta nel mettere in evidenza come il sistema visivo organizzi le forme in parti, un tipo di organizzazione molto utile per il compito del riconoscimento di un oggetto dalla sua forma. La superficie cosinusoidale ci mostra anche che rovesciando una forma la sua organizzazione può cambiare. Questo vuol dire che il sistema visivo è un po' capriccioso nella sua organizzazione? Sembra improbabile. Se non è governato a capriccio, allora, deve seguire delle regole nella definizione delle parti. E se le regole non debbono essere arbitrarie, debbono essere fondate su qualche legge o regolarità nel mondo esterno.

Questa linea di ragionamento ha portato Whitman A. Richards del MIT e me a cercare una legge o una regolarità che possa motivare un insieme di regole per la suddivisione in parti delle superfici. Abbiamo scoperto che quella pertinente è la regolarità di trasversalità, che può essere formulata in questo modo: quando due superfici di forma arbitraria possono penetrare casualmente l'una nell'altra, si incontrano sempre a un contorno di discontinuità concava dei loro piani tangenti (*si veda l'illustrazione in alto della pagina a fronte*). Da questa formulazione la regolarità di trasversalità può suonare un po' esoterica, ma in effetti è una parte familiare dell'esperienza quotidiana. Una cannuccia in una bibita, per esempio, forma una discontinuità concava circolare là dove incontra la superficie della bibita. Un candito in un dolce, le punte di una forchetta in una bistecca, una sigaretta in bocca: sono tutti esempi di questa onnipresente regolarità.

In base alla regolarità della trasversalità si può proporre una prima regola per la partizione di una superficie: si divida una superficie in parti lungo tutti i contorni di discontinuità concava. Questa regolarità non ci può aiutare con la superficie cosinusoidale, perché è perfettamente continua. Prima bisogna generalizzare in qualche modo la regola, come faremo più avanti. Nella sua forma non generalizzata, però, la regola può già spiegare parecchie ben note situazioni percettive.

Per esempio, dalla regola si trae l'ovvia previsione che le parti della scala raffigurata nell'illustrazione centrale di pagina 106 sono i suoi scalini, e che ogni scalino giace fra due linee successive di discontinuità concava nella scala. Dalla regola si trae anche una previsione meno ovvia. Se la scala subisce una inversione percettiva, così che la «figura» diventi «sfondo» e viceversa, allora i confini fra gli scalini debbono mutare. Questa conclusione segue perché solo le discontinuità concave definiscono i confini fra gli scalini, e quella che appare come una concavità da un lato di una superficie deve apparire come una convessità dall'altro lato. Pertanto, quando la scala subisce un'inversione, le discontinuità convesse e concave debbono invertirsi i loro ruoli, dando così nuovi confini per gli scalini. Potete mettere alla prova da voi questa previsione osservando lo scalino le cui due facce sono in colore. Quando la scala subisce un'inversione, noterete che le due strisce di colore non sono più su un solo scalino, bensì su due scalini adiacenti.

Questa previsione può essere confermata con una dimostrazione più complessa, per esempio con il test dei cubi impilati visibile nella stessa illustrazione. Le tre facce in colore, che all'inizio appaiono far parte di un unico cubo, quando la figura subisce un'inversione vengono viste come facce di tre cubi diversi.

Da questa semplice regola di partizione segue un'ulteriore previsione. Se la regola non definisce un'unica partizione di qualche superficie, allora il modo appropriato di dividere in parti quella superficie dovrebbe essere percettivamente ambiguo (a meno che esistano regole ulteriori in grado di eliminare l'ambiguità). Si può avere una chiara conferma di questo fatto facendo riferimento al blocco a forma di

Le linee di curvatura sono facilmente identificate su un bicchiere di forma cilindrica. Le linee di massima curvatura (*a sinistra*) sono cerchi; le linee di minima curvatura (*a destra*) sono rette.

I confini delle parti, come definiti dalla regola di partizione generalizzata per superfici continue, sono rappresentati dai contorni in colore su questa superficie di forma arbitraria. Le linee nere sono linee di massima curvatura i cui minimi danno luogo ai contorni di suddivisione in colore.

Questa curva piana che subisce una inversione, costruita da Fred Attneave dell'Università dell'Oregon tracciando a caso una linea a metà di un cerchio e separando poi le due metà, mostra che la forma apparente del contorno dipende da quale è il lato della linea percepito come figura.

Si può ottenere una figura che mostra un'analoga inversione con una curva piana non continua. Si può vedere il contorno a zig-zag come una catena di montagne, alternativamente alte e basse, oppure, con l'inversione fra figura e sfondo, come una catena di montagne alte con picchi accoppiati.

L nell'illustrazione in basso di pagina 106. L'unica discontinuità concava è la linea verticale nell'incavo della *L*. Di conseguenza la regola non definisce un'unica partizione del blocco. In senso percettivo, vi sono tre modi plausibili di tagliare il blocco in parti e tutti e tre si basano sul contorno definito dalla regola di partizione, ma vengono completati lungo percorsi diversi.

Nonostante la sua semplicità questa regola di partizione conduce a idee interessanti sulla percezione della forma. Per studiare la superficie cosinusoidale e altre superfici continue, però, la regola deve essere generalizzata. Per questo dobbiamo fare una breve digressione sulla geometria differenziale delle superfici, per chiarire tre concetti importanti: quello di normale a una superficie, quello di curvatura principale e quello di linea di curvatura. Sono concetti molto tecnici ma, per fortuna, è possibile darne facilmente una caratterizzazione intuitiva.

La normale a una superficie in un suo punto può essere pensata come un ago di lunghezza unitaria che emerge perpendicolarmente dalla superficie in quel punto, simile agli aculei di un riccio di mare. Collettivamente le normali a tutti i punti di una superficie prendono il nome di «campo di normali alla superficie». Di solito vi sono due possibili campi di normali alla superficie, per una superficie data: le normali possono essere dirette verso l'esterno oppure verso l'interno. Per esempio, una sfera può avere tutte le sue normali alla superficie dirette radialmente verso l'esterno come aculei, o tutte dirette verso l'interno, ossia verso il suo centro. Adottiamo la convenzione che il campo delle normali alla superficie sia scelto sempre in modo che le normali siano orientate verso l'interno di ciò che costituisce «figura». Così una palla da baseball ha le normali dirette verso l'interno, mentre una bolla sott'acqua ha normali dirette verso l'esterno. Invertendo la scelta di figura e sfondo su una superficie si determina al contempo un cambiamento nell'orientazione delle normali alla superficie. Un'inversione del campo delle normali alla superficie induce un cambiamento nel segno di tutte le curvature principali in ogni punto della superficie.

Spesso è importante conoscere non solo la normale alla superficie in un punto, ma anche il modo in cui, in quel punto, la superficie si curva. Il matematico svizzero Leonhard Euler, nel XVIII secolo, scoprì che per ogni punto di qualunque superficie vi è sempre una direzione in cui la curvatura della superficie è minima e una seconda direzione, sempre ad angolo retto rispetto alla prima, lungo la quale la curvatura della superficie è massima. (Nel caso di un piano o di una sfera la curvatura della superficie è identica in tutte le direzioni in ogni punto.) Queste due direzioni sono chiamate direzioni principali, e le corrispondenti curvature della superficie sono chiamate curvature principali. Partendo in un certo punto e spostandosi sempre lungo la direzione di massima curvatura principale, si traccia una linea di curvatura massima. Spostandosi invece nella direzione di minima curvatura principale si traccia una linea di curvatura minima. Su un comune bicchiere da tavola la famiglia delle linee di curvatura massima è un insieme di cerchi attorno al bicchiere; le linee di curvatura minima sono linee rette che vanno dall'orlo alla base del bicchiere (*si veda l'illustrazione in alto nella pagina precedente*).

Avendo presenti questi concetti, la regolarità della trasversalità può essere estesa facilmente a superfici continue. Supponiamo che, ogniqualvolta una superficie possiede una discontinuità concava, si elimini in qualche modo la discontinuità, magari stendendo al di sopra di essa una pellicola. Allora una discontinuità concava diventa, detto in termini intuitivi, un'isolinea lungo la quale la superficie presenta localmente la massima curvatura negativa. Più precisamente, la versione generalizzata della trasversalità suggerisce la seguente generalizzazione della regola di partizione di una superficie: si divida una superficie in parti in corrispondenza dei minimi negativi di ciascuna curvatura principale lungo la famiglia di linee di curvatura a essa associata (*si veda l'illustrazione in basso nella pagina precedente*).

Questa regola suddivide la superficie cosinusoidale lungo le isolinee circolari in colore e spiega anche perché le parti siano diverse quando la pagina viene rovesciata: il sistema visivo allora effettua una inversione nel suo modo di assegnare le categorie di figura e sfondo sulla superficie (forse in grazia di una preferenza per un'interpretazione che pone l'oggetto al di sotto del punto di vista dell'osservatore, anziché al di sopra). Quando figura e sfondo si invertono, altrettando accade al campo delle normali alla superficie, in accordo con la convenzione citata in precedenza. Tuttavia con qualche calcolo semplice si vede che, quando si invertono le normali, si inverte anche il segno delle curvature principali. Di conseguenza i minimi delle curvature principali debbono diventare massimi e viceversa. Poiché i minimi delle curvature principali sono utilizzati per definire i confini delle parti, ne segue che anche questi confini debbono spostarsi. Per riassumere, le parti cambiano perché la regola di partizione, motivata dalla regolarità della trasversalità, sfrutta i minimi delle curvature principali, e questi minimi vengono spostati sulla superficie quando si verifica una inversione di figura e sfondo.

La regolarità della trasversalità, in breve, offre una unità di fondo per le spiegazioni della percezione delle parti nel caso sia di superfici continue, sia di superfici discontinue. Tale regolarità è alla base della spiegazione anche di un'altra classe ben nota di illusioni visive: quella delle curve nel piano che presentano inversione. Un buon esempio di questo fenomeno ci è fornito dalla figura escogitata da Fred Attneave dell'Università dell'Oregon (*si veda l'illustrazione in alto nella pagina a fronte*). Attneave ha trovato che basta tracciare a caso una linea a metà di un cerchio e separare le due metà per creare due contorni dall'aspetto molto diverso. Evidentemente, come sottolinea Attneave stesso, il modo in cui il contorno appare dipende dal lato che viene assunto come parte della figura, e non da una qualche precedente familiarità con il contorno stesso (si veda l'articolo *La multistabilità nella percezione* di Fred Attneave in «Le Scienze», n. 43, marzo 1972).

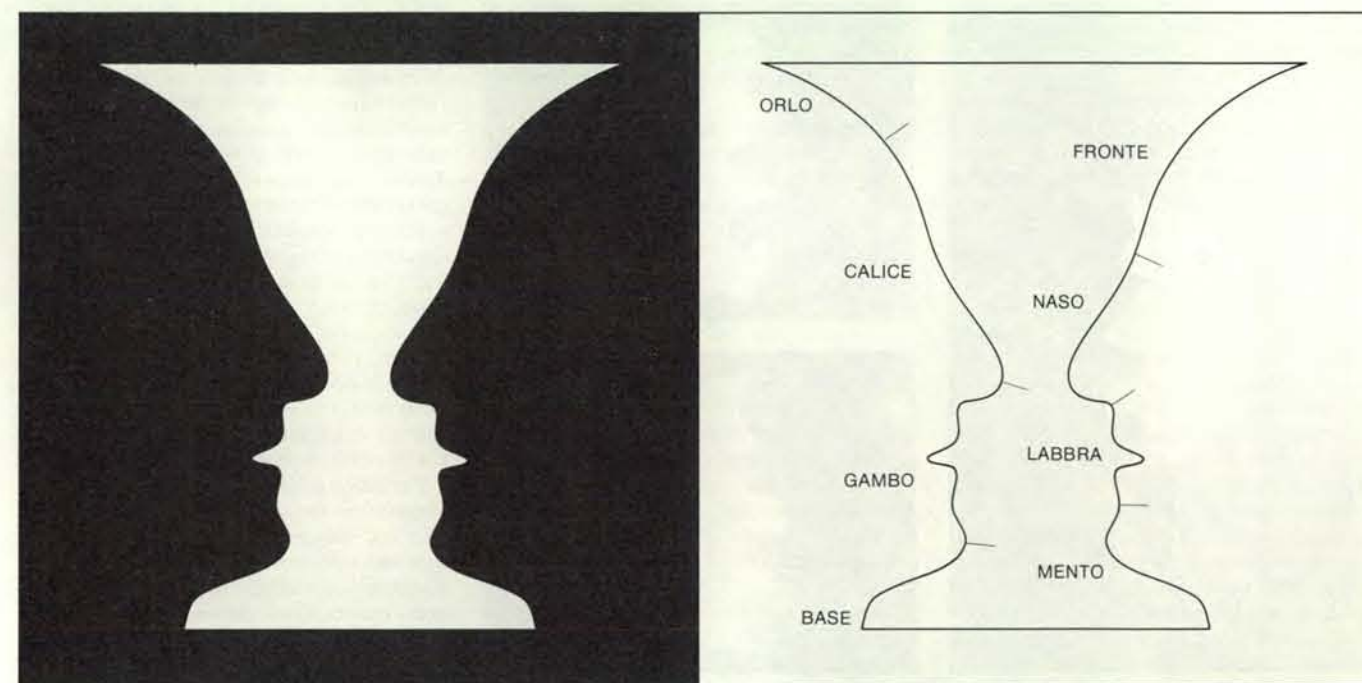
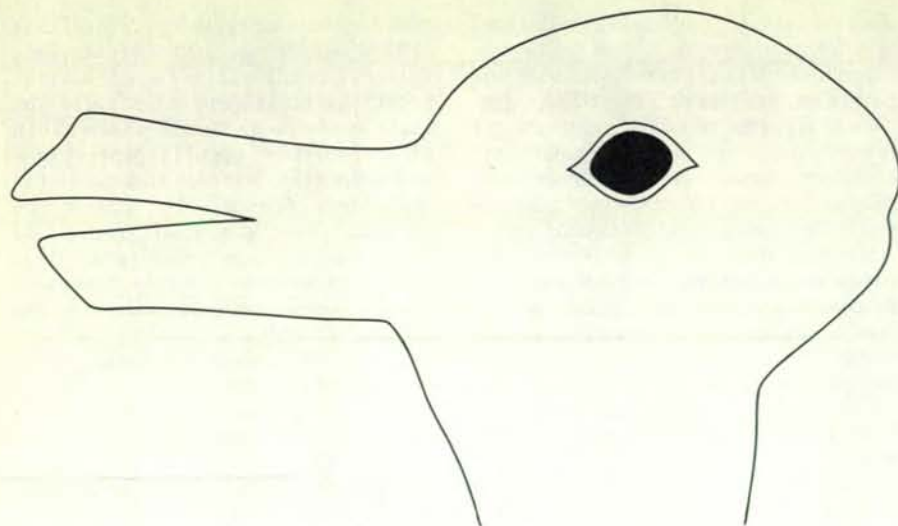
In che modo questo fenomeno viene spiegato grazie alla regolarità della trasversalità? La risposta comporta tre passi: (1) una proiezione della regolarità della trasversalità da tre a due dimensioni; (2) una breve digressione sulla geometria differenziale delle curve nel piano; e (3) la formulazione di una regola di partizione per le curve nel piano.

La versione della regolarità della trasversalità per le due dimensioni è simile a quella per le tre dimensioni. Se due superfici di forma arbitraria possono penetrare l'una nell'altra in modo casuale, allora in qualunque proiezione bidimensionale della loro superficie composita si incontreranno sempre in cuspidi concave. Per dirla in parole più povere, in una silhouette si formano sempre cuspidi concave in corrispondenza dei punti dove finisce una parte e ne comincia un'altra. Il che suggerisce la seguente regola di partizione per le curve nel piano: si divida in parti una curva nel piano in corrispondenza delle cuspidi concave. Questa regola non può essere applicata alla dimostrazione di Attneave, perché la sua dimostrazione è basata su un contorno ovunque continuo. Ancora una volta, è necessario generalizzare la regola. Ciononostante, nella sua forma non generalizzata la regola può spiegare una versione della figura di Attneave che non è ovunque continua.

Nell'illustrazione inferiore della pagina a fronte lo stesso contorno a zig-zag può sembrare una catena in cui si alternano montagne più o meno alte o, per l'assegnazione inversa di figura e sfondo, una catena di montagne alte, con picchi accoppiati. Il contorno è organizzato in parti in modo diverso quando figura e sfondo si invertono, perché la regola di partizione usa, per definire i confini fra le parti, solo cuspidi concave. Quella che è una cuspide concava, se un lato del contorno è figura, deve diventare una cuspide convessa quando è figura l'altro lato, e viceversa. C'è un parallelismo fra questo esempio e la scala di Schröder esaminata in precedenza.

Prima di generalizzare la regola ai contorni continui, vediamo in breve due concetti della geometria differenziale delle curve nel piano: quello di normale principale e quello di curvatura. La normale principale a un punto di una curva può essere pensata come un ago di lunghezza unitaria che emerge perpendicolarmente alla curva in tale punto. Tutte le normali principali a tutti i punti di una curva formano un campo di normali principali. Di solito vi sono due possibili campi di normali principali, uno per ciascun «lato» della curva. Adottiamo la convenzione che il campo delle normali principali sia scelto sempre orientato verso il lato della curva che costituisce la figura. Quando si inverte la scelta fra figura e sfondo su una curva si determina al contempo un cambiamento nel campo delle normali principali. Quel che ci importa di notare è che, a causa della convenzione che vuole le normali principali orientate verso l'interno della figura, le parti concave di una curva continua hanno curvatura negativa e le parti convesse curvatura positiva.

È facile, ora, generalizzare la regola di partizione per le curve nel piano. Supponiamo che, ogniqualvolta una curva ha una cuspide concava, sia possibile smussare un po' la curva stessa. Allora una cuspide concava diventa un punto di curvatura negativa che, localmente, presenta il massimo valore assoluto di curvatura. Questo ci porta alla seguente regola generalizzata di partizione: si divida in parti una curva nel piano in corrispondenza dei minimi negativi di curvatura.

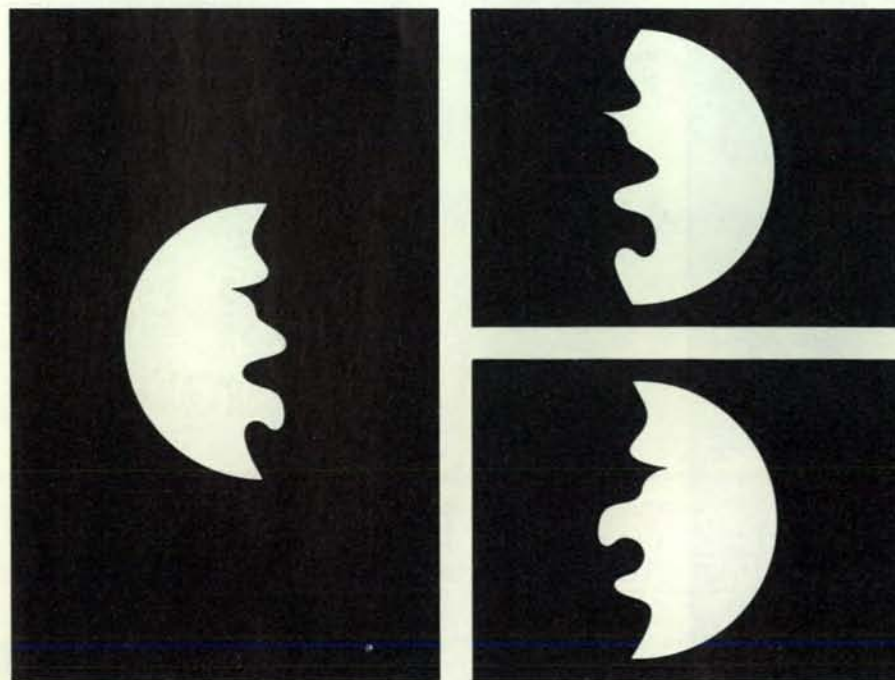
Ora è possibile spiegare perché le due metà del cerchio di Attneave sembrano così diverse. Quando figura e sfondo si invertono, si inverte anche il campo delle normali principali, in accordo con la convenzione, e quando le normali princi-

L'illusione delle due facce e del vaso, trovata da Edgar Rubin intorno al 1915: la figura può essere vista come una coppia di profili di facce umane o come un vaso (*a sinistra*). Se si assume come figura una faccia, la suddivisione in parti della figura con riferimento ai minimi di curvatura divide il contorno in blocchi che corrispondono a una fronte, un naso, un paio di labbra e un mento; se si assume come figura il vaso, invece, la definizione dei confini tra le parti mediante i minimi di curvatura divide il contorno in un orlo, un calice, un gambo e una base (*a destra*). In ambedue i casi le linee normali principali (rappresentate dai filetti) sono orientate verso l'interno di quello che costituisce la «figura».

L'illusione dell'animale che subisce un'inversione non comporta un'inversione di figura e sfondo. Corrispondentemente, i confini fra le parti definiti dai minimi di curvatura non cambiano di posizione quando muta l'interpretazione. Le orecchie del coniglio diventano il becco dell'anatra.

pali si invertono la curvatura in ogni punto della curva deve cambiare segno. In particolare, i minimi di curvatura devono diventare massimi, e viceversa. Questa ridistribuzione dei minimi di curvatura porta a una nuova suddivisione della curva secondo la regola di partizione. In breve, la curva appare diversa perché è organizzata in blocchi, ovvero unità, fondamentalmente differenti. Si noti che se si scegliesse di definire i confini fra le parti per mezzo dei punti di flesso, cioè sia dei massimi, sia dei minimi di curvatura, i blocchi non cambierebbero in caso di inversione fra figura e sfondo.

Un esempio chiaro di due suddivisioni in parti molto diverse per una curva si può avere con la famosa illusione del vaso e dei due volti, trovata da Edgar Rubin intorno al 1915 (*si veda l'illustrazione nella pagina precedente*). Se si assume come figura una faccia, i minimi di curvatura dividono la curva in blocchi che corrispondono alla fronte, al naso, al labbro superiore, al labbro inferiore e al mento. Se si assume come figura, invece, il vaso, i minimi vengono ridistribuiti e dividono la curva in nuovi blocchi che corrispondono a una base, una coppia di parti del gambo, un calice e un bordo del calice. Probabilmente non è un caso che le parti definite dai minimi abbiano spesso un nome nel vocabolario quotidiano.

Sono state trovate presentazioni visive che, come il vaso di Rubin, ammettono più di una interpretazione per un solo contorno ma non comportano un'inversione fra figura e sfondo. Un esempio famoso è quello dell'illusione coniglio-anatra (*si veda l'illustrazione in alto in questa pagina*). Poiché tali illusioni non comportano una inversione fra figura e sfondo, e poiché i minimi di curvatura non cambiano la loro posizione, la regola di partizione deve prevedere che i confini delle parti siano identici per ambedue le interpretazioni di ciascun contorno. Questa previsione trova facilmente conferma. Le orecchie del coniglio, per esempio, diventano parte del becco dell'anatra.

Se il sistema visivo umano segue veramente la regola dei minimi per la partizione delle curve, ci si aspetterebbe che la regola possa prevedere alcuni giudizi di somiglianza di forma. Talvolta la previsione può essere controintuitiva: un esempio di previsione di questo genere si può vedere nell'illustrazione in basso di questa pagina. Guardate brevemente la mezza luna sulla sinistra dell'illustrazione; poi guardate rapidamente le due mezze lune a destra e stabilite quale vi sembra più simile alla prima. In un esperimento effettuato con numerose figure simili, Aaron F. Bobick del MIT e io abbiamo trovato che quasi tutti i soggetti scelgono la mezza luna superiore è identico a quello della mezza luna di sinistra, a meno della mezzaluna superiore è identico a quello della mezza luna di sinistra, a meno dell'inversione fra figura e sfondo. Il contorno della mezza luna inferiore è stato ottenuto per inversione speculare e due parti definite dai minimi di curvatura sono scomparse. Perché però quella inferiore continua a sembrare più simile a quella di sinistra? La regola dei minimi ci dà una risposta molto semplice. Il contorno inferiore, che non è stato ottenuto dal contorno originale per inversione fra figura e sfondo, possiede gli stessi confini delle parti. Il contorno in alto, che invece è stato ottenuto per inversione fra figura e sfondo dal contorno originale, presenta confini diversi fra le parti.

Per riassumere, la visione è un processo attivo la cui funzione è inferire descrizioni utili del mondo dalle mutevoli figure di luce che cadono sulla retina. Queste descrizioni sono affidabili solo nella misura in cui i processi di inferenza che le costruiscono sfruttano regolarità nel mondo visivo, come la rigidità, la planarità e la trasversalità. La scoperta di regolarità rilevanti e lo studio matematico della loro potenza nel guidare le inferenze visive prospettano promettenti direzioni di ricerca per tutti coloro che si sforzano di capire la visione umana.

Il test della mezza luna dimostra che i giudizi sulla somiglianza di forme possono essere previsti correttamente sulla base della regola di partizione fondata sui minimi di curvatura. A prima vista la mezza luna a destra in basso sembra più simile alla mezza luna a sinistra della mezza luna in alto a destra. A un esame più attento, però, si nota che il contorno della mezza luna in alto a destra è identico a quello della mezza luna a sinistra, mentre quello della mezza luna in basso a destra è stato ottenuto per inversione speculare e inoltre presenta due parti scambiate fra di loro.

(RI)CREAZIONI AL CALCOLATORE

di Brian Hayes

Dove si parla dell'automa finito: un modello minimale delle trappole per topi, dei ribosomi e dell'anima umana

I calcolatori più potenti non hanno né hardware né software: sono fatti di puro pensiero. La più celebre tra queste macchine astratte è quella ideata nel 1936 dal matematico inglese Alan Mathison Turing. È una macchina in grado di fare più di quanto abbia mai potuto fare un calcolatore costruito di silicio: in realtà può calcolare tutto ciò che è possibile calcolare. Esiste poi una classe di calcolatori concettuali che, pur non raggiungendo l'onnipotenza della macchina di Turing, sono altrettanto interessanti. Si tratta delle cosiddette macchine finite, o automi finiti, che determinano le caratteristiche minime di un calcolatore digitale funzionante.

Una corretta definizione di macchina finita richiederebbe un grado di rigore matematico inadeguato per questa sede. Si può però chiarire la natura del concetto con qualche esempio. Cercando qualche macchina finita, ne ho trovato un ottimo esempio in una stazione della metropolitana sulla Lexington Avenue di New York. È un cancelletto girevole di vecchio tipo, fatto non con il consueto treppiede compatto d'acciaio, ma di quattro bracci incrociati di quercia, consumati da innumerevoli mani e fianchi.

Il cancelletto ha due stati: bloccato e sbloccato. Supponiamo che si trovi nello stato bloccato, in modo che non si possano ruotare i bracci. Inserendo un gettone si ottiene una certa modificazione del meccanismo interno che consente ai bracci di muoversi; in altre parole, il gettone induce una transizione allo stato sbloccato. Ruotando i bracci di 90 gradi si provoca un'altra transizione che riporta il cancelletto allo stato di blocco. Nella figura della pagina a fronte si vedono le transizioni rappresentate in modo schematico. Gli stati del sistema sono rappresentati da nodi (riquadri) e le transizioni da archi (frecce) che li collegano.

Nell'analisi finita del cancelletto, l'inserzione di un gettone e la rotazione dei bracci sono i possibili input del sistema. La risposta della macchina dipende sia dall'input sia dallo stato al momento dell'input. Spingere i bracci quando il cancelletto non ha ricevuto un gettone non vi garantirà un viaggio in metropolitana. Inserire un gettone quando i bracci sono già sbloccati è altrettanto inutile, anche se in un modo leggermente diverso. Il secondo gettone viene accettato ma non ha effetto sullo stato della macchina: può passare una sola persona e poi il cancelletto torna a bloccarsi. La macchina può accettare anche tre o quattro gettoni di seguito, ma uno solo ha effetto. Forse qualche scettico vorrebbe altre prove prima di accettare la generalizzazione secondo cui tutti i gettoni dopo il primo non hanno effetto; in tal caso, però, dovrà vedersela con i suoi gettoni.

La ragione per cui il cancelletto non può dare più passaggi per più gettoni è che non ha modo di contare i gettoni che riceve. La sua sola forma di memoria è del tutto rudimentale: passando da uno stato all'altro «ricorda» se l'input più recente era un gettone o una spinta sui bracci. Tutti gli input precedenti vanno perduti. Val la pena di notare che questa smemoratezza non va mai a svantaggio della città. Le cose potrebbero andare ben peggio: si potrebbe progettare un cancelletto che cambi di stato dopo ogni moneta, indipendentemente dallo stato attuale, nel quale caso due gettoni in fila non farebbero passare nessuno.

Il cancelletto illustra la maggior parte delle proprietà essenziali di una macchina finita. Ovviamente la macchina deve avere degli stati, che possono essere solo in numero finito. Ci possono essere input e output associati a ogni stato. Gli stati devono essere discreti, cioè chiaramente distinguibili, e le transizioni da uno stato all'altro devono essere efficacemente istantanee. Molto dipende dal punto di vista: giorno e notte sono stati discreti se si vogliono definire l'alba e il tramonto come processi istantanei. La macchina è fatta solo dell'insieme degli stati, degli input e degli output; non ci possono essere dispositivi ausiliari, e in particolare non può esserci alcuna possibilità d'immagazzinamento di informazioni.

Le regole per la costruzione di un automa finito consentono delle varianti. Ci sono automi deterministici e non deterministici, automi di Moore e automi di Mealy. In un automa deterministico, un dato input in un dato stato produce invariabilmente lo stesso risultato; in un automa non deterministico ci possono essere più transizioni possibili. Nell'automa di Moore (dal nome di Edward F. Moore) ogni stato ha un unico output; nell'automa di Mealy (dal nome di G. H. Mealy) gli output sono associati alle transizioni invece che agli stati. Risulta, però, che la varietà di architetture è un po' un'illusione: qualsiasi compito possa essere eseguito da un certo tipo di macchina finita può essere eseguito anche dagli altri tipi, anche se può variare il numero di stati necessari. In questa sede parlerò soprattutto degli automi di Moore deterministici, quelli con la struttura più semplice.

Se vi mettete alla ricerca di macchine finite, ne troverete dappertutto. I congegni a moneta sono gli esempi favoriti dei manuali. Alcune macchine emettitrici sono meno rapaci del cancelletto della metropolitana: una volta ricevuta la somma giusta, entrano in uno stato in cui tutte le monete ulteriori sono respinte. Il congegno a moneta con il maggior numero di stati possibile è sicuramente la *slot machine* di Las Vegas. In linea di principio è deterministica, nondimeno è decisamente difficile trovare un input (una moneta e una pressione sulla leva) che provochi una transizione a un particolare stato finale.

Molti apparecchi domestici possono essere visti come automi finiti, sia pure particolarmente ottusi. Una lavabiancheria passa attraverso un'inflessibile successione di stati - riempimento, lavaggio, risciacquo, centrifuga - e i pochi input dotati di significato, come il togliere la spina dalla presa elettrica, hanno di solito lo stesso effetto in tutti gli stati. In modo analogo, un semaforo ha un piccolo repertorio di stati che si ripetono indefinitamente. Il più assillante di tutti gli automi finiti è a mio giudizio un orologio digitale. Se visualizza il mese, il giorno e il passaggio delle ore, dei minuti e dei secondi, ha qualcosa come 31 milioni di stati e nel corso di un anno visita ogni stato esattamente una volta.

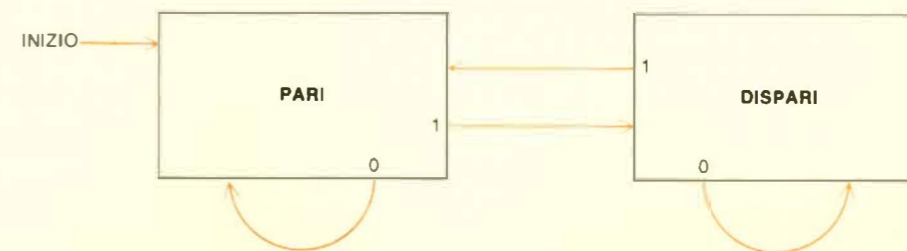
Una trappola per topi è un automa finito: il topo, di solito a suo scapito, innesca una transizione dallo stato di carica allo stato di scatto. Una serratura a combinazione è un automa finito con molti input possibili, uno solo dei quali provoca una transizione di stato. Un telefono ha stati che potrebbero essere definiti agganciato, sganciato, attesa, segnale, composizione del numero, squillo, collegato e fuori uso. Un'automobile può dimostrare efficacemente come l'effetto di un input vari a seconda dello stato del sistema. Che cosa succede quando si preme a fondo l'acceleratore? Dipende. La frizione è inserita? Il freno a mano è sbloccato? La marcia è ingranata? È una marcia avanti o la retromarcia? La porta del garage è aperta?

Nella cellula vivente, il sistema molecolare costituito dal ribosoma e dalle varie specie di RNA di trasporto opera come un automa finito. Gli input sono le quattro basi nucleotidiche dell'RNA messaggero, indicate con le abbreviazioni *U*, *A*, *G* e *C*. Gli output sono i 20 amminoacidi che compongono le proteine. Una catena di nucleotidi è riconosciuta come input valido per l'automa solo se inizia con il segnale di «partenza» *AUG*. In seguito l'automa legge in modo continuo il flusso di input, cambiando stato ogni volta che riconosce un codone, ossia una tripletta di nucleotidi. I tre codoni speciali *UAA*, *UAG* e *UGA* sono segnali di «fine»: quando ne incontra uno, l'automa si ferma. Molti altri sistemi biologici possono essere rappresentati come automi finiti; esempi che vengono in mente sono la molecola di emoglobina e le proteine promotore e repressore dei batteri.

Nella teologia di Tommaso d'Aquino l'anima è un automa finito, meravigliosamente elaborato e totalmente deterministico. È creata in uno stato di rischio, come conseguenza del peccato originale. Con il battesimo entra in uno stato di grazia, ma certi atti (idolatria, bestemmia, adulterio e così via) inducono una transizione a uno stato di peccato. Sono allora necessari confessione, pentimento e assoluzione per riportare l'anima allo stato di grazia. L'effetto di un input finale, la morte, dipende tutto dallo stato dell'anima al momento della morte: in uno stato di grazia la morte porta alla salvezza ma in uno stato di peccato porta alla dannazione. La macchina anima è in realtà più complessa di quanto faccia pensare questa descrizione: bisognerebbe distinguere tra i vari gradi di peccato (veniale e capitale, attuale e abituale) e si dovrebbe tener conto di altri possibili stati dell'anima (come quelli associati al limbo e al purgatorio) e di altri possibili input (come il Giudizio Universale).

Nella meccanica quantistica anche l'atomo diviene un automa finito, quindi lo stesso avviene per ogni cosa fatta di atomi. Gli stati dell'atomo sono i livelli di energia consentiti; gli input e gli output sono i fotoni, quanti di radiazione elettromagnetica. In una descrizione accurata, penso che l'atomo sarebbe classificato come automa di Mealy non deterministico con transizioni epsilon. È una macchina non deterministica perché l'effetto di un input non può essere previsto con certezza. È un automa di Mealy perché la natura dell'output (vale a dire l'energia del fotone) è determinata dalla transizione, non dallo stato in cui la macchina è entrata. Le transizioni epsilon sono quelle che possono aver luogo in assenza di qualsiasi input; devono essere incluse nel modello perché un atomo può emettere un fotone e cambiare di stato del tutto spontaneamente.

Il cervello è un automa finito? Per coincidenza, i moderni studi sui sistemi finiti iniziarono proprio con un modello di reti di neuroni ideato nel 1943 da Warren S. McCulloch e Walter Pitts. I neuroni di McCulloch e Pitts erano semplici cellule con input eccitanti e inibenti; ogni cellula aveva un solo output e due stati interni: eccitato e non eccitato. Le cellule potevano essere disposte in reti per compiere

Un diagramma delle transizioni di stato per un cancelletto di metropolitana

La macchina per il controllo della parità

varie funzioni logiche, tra cui le funzioni «and», «or» e «not» che sono ora elementi di uso comune nei sistemi logici elettronici. Stephen C. Kleene, dell'Università del Wisconsin a Madison, dimostrò nel 1956 l'equivalenza tra le reti di neuroni ideali e i diagrammi delle transizioni di stato che abbiamo illustrato qui.

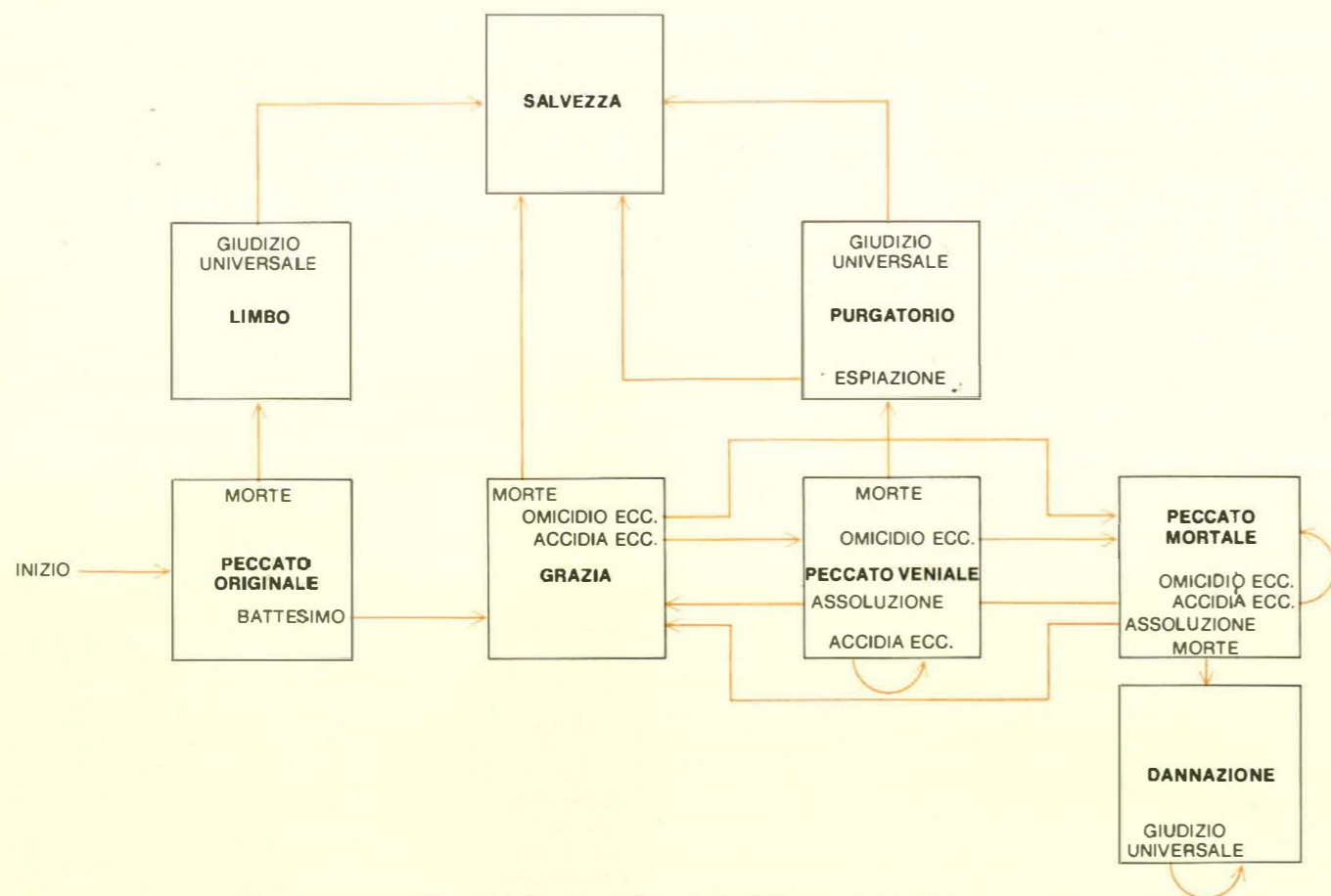
Quarant'anni dopo il lavoro di McCulloch e Pitts, è ancora argomento di discussione la possibilità di classificare il cervello tra i sistemi finiti. Naturalmente, il numero di neuroni è necessariamente finito, ma questo non è il solo problema. Un vero neurone è molto più complesso di una cellula a due stati e alcune delle sue proprietà possono variare con continuità, anziché essere vincolate a occupare stati discreti. Inoltre, il divieto di un magazzino ausiliario di informazioni in un modello a stati finiti è quanto mai imbarazzante. Se la vita mentale non è altro che una successione di stati istantanei, senza conoscenza della propria storia, allora che cos'è la memoria?

Gli stati della mente di cui si occupa la psicologia, quali noia, paura, desiderio, gioia e dolore, sembrano rientrare più facilmente nel complesso di una teoria a stati finiti. D'altra parte, gli stati sono così numerosi e così limitata è la comprensione delle transizioni, che il modello sembra inutilizzabile. Solo per animali di minor complessità è possibile tracciare più di qualche isolato frammento del diagramma delle transizioni di stato e in quelle specie lo sperimentatore può non avere alcun accesso diretto ai presunti stati mentali. In effetti, questa linea di ricerca è stata seguita soprattutto dai comportamentisti, che negano l'effettiva esistenza di stati mentali.

Anche il caso del calcolatore digitale - e intendo qui la macchina tangibile, l'hardware - è problematico. Il comune modello mentale di un calcolatore, formulato da John von Neumann, divide la macchina in una unità centrale di elaborazione e una schiera di celle di memoria. Il concetto di stati finiti è indubbiamente applicabile alle varie componenti dell'unità centrale di elaborazione quali registri, sommatori e il meccanismo di controllo deputato a dirigere le operazioni interne dell'unità di elaborazione.

I guai iniziano quando si prende in considerazione la memoria. Secondo le regole per la costruzione di un automa finito, non è consentita alcuna memoria esterna, quindi ogni cella deve essere vista non come un elemento di immagazzinamento separato dall'unità di elaborazione, ma come una parte dello stato complessivo della macchina. Se tutte le celle sono vuote, il calcolatore è in uno stato; se si riempie una sola cella, entra in un altro stato, e così via. Questa concezione del calcolatore è ben poco illuminante, in parte perché non stabilisce alcun collegamento tra lo stato della macchina e ciò che essa sta facendo. Inoltre, il numero di stati è immenso. Forse nemmeno un calcolatore di modeste dimensioni (100 elementi binari), che avesse girato continuamente per tutta l'età dell'universo, sarebbe passato per tutti i propri stati.

Il ruolo fondamentale dell'automa finito nella scienza del calcolatore è a un livello d'astrazione maggiore dei meccanismi da orologeria dell'hardware. Un calcolatore che giri sotto la direzione di un programma non è più un insieme di porte logiche, registri, celle di memoria e altri aggeggi elettronici; è una macchina «virtuale» le cui parti funzionanti sono definite dal programma e possono essere ridefinite se necessario. Mentre l'hardware conosce solo gli interi binari e semplici istruzioni per trasferirli e manipolarli, il calcolatore virtuale ha a che fare con sistemi simbolici molto più espressivi: parole, equazioni, matrici, funzioni, vettori, codoni, liste, immagini, magari addirittu-

Gli stati dell'anima nella teologia di Tommaso d'Aquino

ra idee. Le tecniche a stati finiti possono essere di qualche validità per la creazione del calcolatore virtuale, e a volte il calcolatore virtuale *è* un automa finito.

Consideriamo un programma il cui obiettivo sia leggere una serie di cifre binarie (1 o 0) e riferire se il numero di 1 ricevuti sia pari a dispari. (Questo compito ha un significato pratico; programmi di controllo della parità sono impiegati, per esempio, nell'individuazione di errori nella trasmissione telefonica di dati digitali.) Il programma può essere costruito come un automa finito con due stati, come si vede nella figura in basso della pagina precedente. L'operazione inizia nello stato pari perché nessun 1 è stato ricevuto e 0 è considerato un numero pari. Ogni 1 nel flusso di input provoca un cambiamento di stato, mentre uno 0 ricevuto nell'uno o nell'altro stato lascia immodificato lo stato stesso. Anche se la macchina non può «ricordare» alcun input che preceda quello più recente e certamente non può contare gli 1 o gli 0, il suo output riflette sempre la parità del flusso di input.

Il modello a stati finiti del calcolo si ritrova comunemente nei programmi che hanno in qualche modo a che fare con testi o altre informazioni sotto forma linguistica. L'esempio più rilevante si trova nei compilatori: programmi che traducono enunciati di programmazione formulati in un linguaggio sorgente in enunciati equivalenti formulati in un linguaggio oggetto, per lo più il «linguaggio macchina» di un particolare calcolatore. I compilatori e altri programmi di traduzione sono essenziali alla nozione di macchina virtuale, in quanto mediano tra simboli con un significato per l'uomo e quelli riconosciuti dal calcolatore.

La parte di compilazione che si può designare come automa finito è detta analizzatore lessicale o *scanner.* Come il cancelletto della metropolitana, è una macchina inghiotti-gettoni. In questo caso, però, i gettoni sono le parole, le unità lessicali fondamentali, del linguaggio. L'analizzatore esamina ogni gruppo di caratteri e stabilisce se si tratta di un vero «gettone», cioè una parola, come una istruzione o un numero; se non lo è, l'analizzatore lo respinge come privo di senso, proprio come il cancelletto respingerebbe un gettone falso.

L'attività di un analizzatore lessicale può essere illustrata da un automa finito impiegato per riconoscere le parole di un semplice linguaggio, anche se di dominio espressivo limitato: le parole sono fatte esclusivamente di numerali romani. Sono accettati, in realtà, solo numerali romani di forma particolare: devono essere in stretta notazione additiva, così che il 9 è rappresentato da VIIII invece che da IX. (Sembra che gli stessi romani impiegassero la notazione additiva e si ritiene che la forma sottrattiva sia stata un'innovazione germanica.)

Nella figura della pagina a fronte si vede un diagramma delle transizioni di stato per la macchina a numerali romani. Il suo alfabeto di simboli di input comprende le lettere M, D, C, L, X, V, I e inoltre il simbolo di spazio. Tutti gli spazi iniziali sono semplicemente ignorati, ma una volta ricevuta la prima lettera il programma compie un'immediata transizione a uno stato identificato (per convenienza) dal nome della lettera. Se la prima lettera è una M, può essere seguita da qualsiasi carattere appartenente all'insieme accettato, inclusa un'altra M. Se il carattere successivo è una D, però, la situazione è diversa. Dallo stato D non è definita alcuna transizione che riporti allo stato M, perché qualsiasi serie di simboli che comprenda DM non può essere una parola ben formata nel linguaggio dei numerali additivi romani. Inoltre, non c'è transizione dallo stato D allo stato D stesso, quindi anche DD è una sequenza esclusa. (La ragione è che i simboli di «mezzo valore» D, L e V non possono essere ripetuti nei numerali romani corretti.)

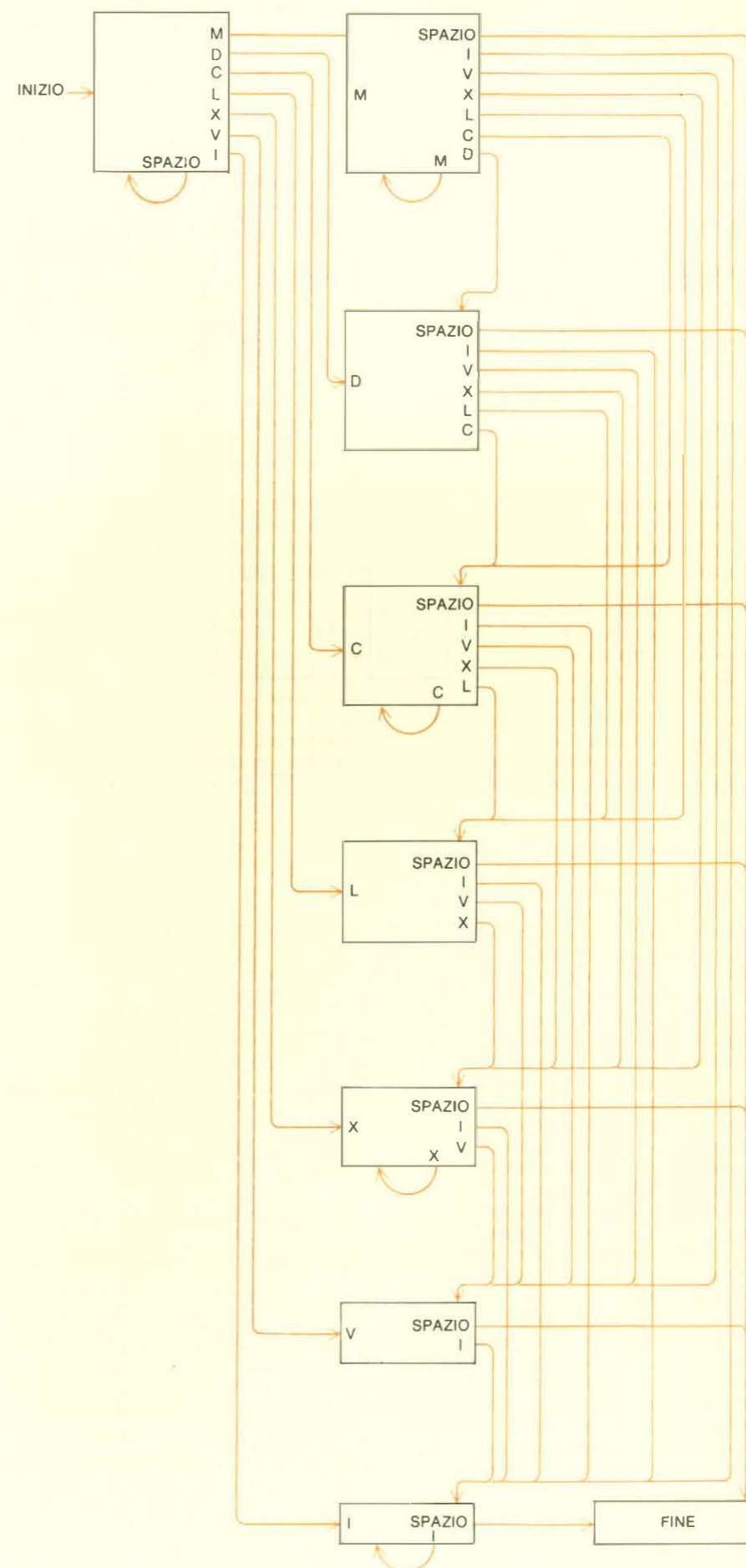
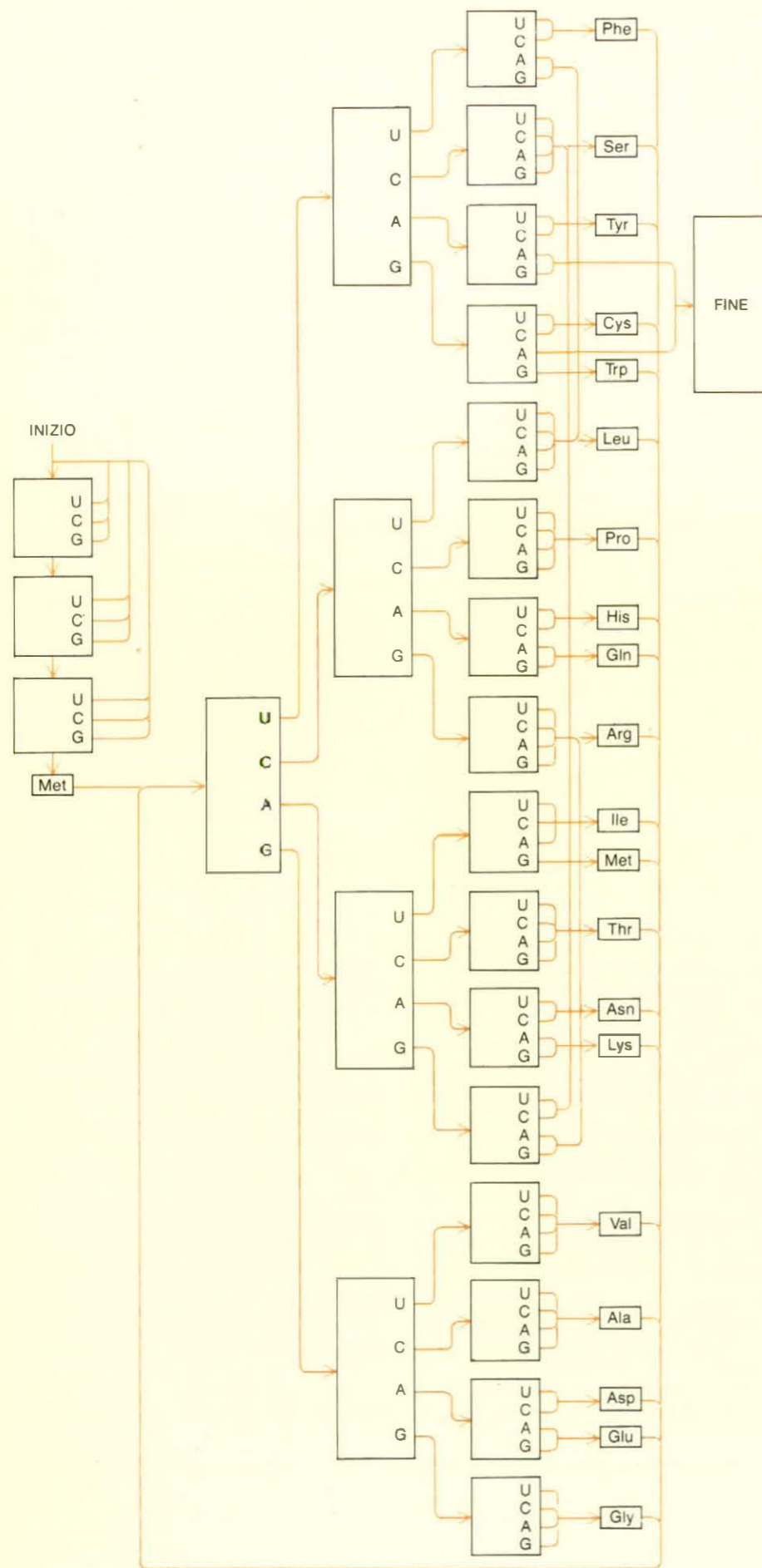
Nello stato D, le uniche lettere accettate sono quelle di valore inferiore: C, L, X, V e I. Lo stesso insieme è accettato nello stato C (perché C può essere ripetuto), ma nello stato L solo le lettere X, V e I sono riconosciute. Dovrebbe essere chiara la regola che governa le transizioni. Gli stati sono disposti in una gerarchia e una volta raggiunto un certo livello la macchina non può mai tornare a un livello superiore; nei livelli di mezzo valore non può mai rimanere allo stesso livello. Quando è raggiunto lo stato I è consentito solo un altro I o uno spazio. Lo spazio, immesso a questo punto o in qualsiasi altro momento dopo la prima lettera, indica la fine della parola e rimanda la macchina allo stato di partenza, pronta a ricevere il successivo numerale romano.

Nessun linguaggio di programmazione a me noto consente l'immissione di numeri in forma romana, ma virtualmente tutti questi linguaggi hanno la possibilità di maneggiare le cifre arabe. Le tecniche per il riconoscimento sono analoghe, anche se c'è una maggiore varietà di formati. Interi semplici come 137 possono essere maneggiati, in linea di principio, da una macchina a uno stato, ma le molteplici parti di un numero come $+6{,}625 \times 10^{-27}$ richiedono un'analisi lessicale più elaborata.

Il sistema ribosoma-RNA di trasporto può essere visto come un analizzatore lessicale che riconosce le sequenze di nucleotidi biologicamente significative in una molecola di RNA messaggero. Per essere accettata, una sequenza deve iniziare con un codone di inizio e terminare con uno dei tre codoni di fine; all'interno di questi confini, è consentita qualsiasi combinazione dei simboli di input *U, A, G* e *C,* presi a tre a tre.

L'analisi lessicale è solo il primo passo nel processo della compilazione. I componenti del compilatore messi in azione dopo il dispositivo di analisi lessicale sono l'analizzatore sintattico o *parser* e il generatore di codice. L'analizzatore sintattico assume come input i gettoni identificati dall'analizzatore lessicale e analizza le loro relazioni sintattiche; è qui che il compilatore si avvicina maggiormente a capire il significato degli enunciati del programma che traduce. Il generatore di codice scrive un programma nel linguaggio oggetto che esegue le funzioni definite dagli enunciati analizzati.

Per i linguaggi giocattolo considerati qui, i compiti dell'analizzatore sintattico e del generatore di codice sono banali. La forma compilata di un enunciato nel linguaggio dei numerali romani potrebbe essere semplicemente l'equivalente arabo del numero e potrebbe essere generata dalla seguente strategia. Prima che una parola sia sottoposta ad analisi lessicale, si specifica una cella di immagazzinamento, che viene posta uguale a zero. Poi, ogni volta che l'analizzatore lessicale entra nello stato M si aggiunge 1000 al valore della cella; per lo stato D si aggiunge 500 e così via. Una volta completata l'analisi lessicale, la cella di memoria contiene il valore del numerale romano. Si noti che il compilatore giocattolo non è più un puro automa finito perché possiede un dispositivo di immagazzinamento ausiliario.

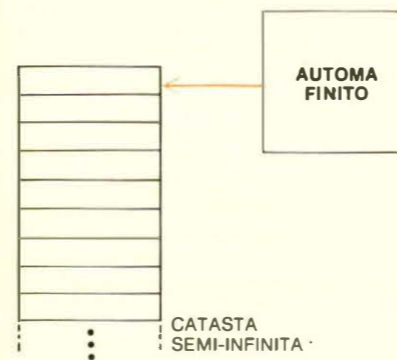
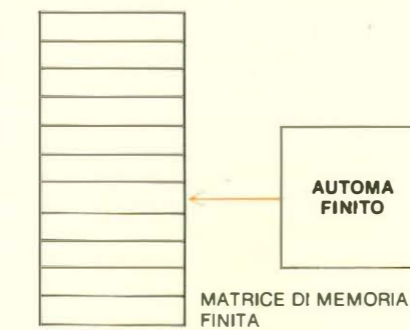
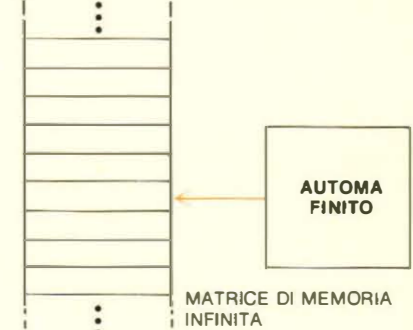
Un compilatore per il codice genetico è ancora più semplice e può essere interamente realizzato nel contesto di un sistema a stati finiti. Il programma «compila-

Un analizzatore lessicale per un linguaggio di numerali romani

Un automa finito traduce il codice genetico in proteina

to» è una sequenza dei simboli di tre lettere standard per gli amminoacidi; i simboli possono essere generati come output degli stati dell'analizzatore lessicale che riconosce i codoni. I tre stati corrispondenti ai codoni di stop non hanno output.

La creazione di un compilatore per un linguaggio abbastanza esteso da essere di utilità generale non è frutto del caso: l'architettura sottostante all'automa finito fornisce almeno un principio di organizzazione. Se la sintassi del linguaggio è specificata con sufficiente precisione, parte del lavoro può anche essere meccanizzata: può essere svolta da un compilatore di compilatore, un programma che ha per input una descrizione formale di un linguaggio e per output un altro programma che traduce enunciati nel linguaggio. Per quanto ne so, nessuno ha ancora pensato a scrivere un compilatore di compilatore di compilatore.

L'identificazione di parole da parte di un analizzatore lessicale è in se stessa un tipo di analisi e l'insieme di tutte le possibili sequenze di simboli in una parola è un tipo di linguaggio. È in realtà un linguaggio infinito: a meno di porre qualche limite artificiale alla lunghezza delle sequenze individuali, si può formare un'infinità varietà di parole riconoscibili. Come fa una macchina con un numero finito di parti a riconoscere un'infinità di enunciati ben formati e a escluderne un'infinità di mal formati? La chiave sta nella struttura del linguaggio stesso. Se gli enunciati di un linguaggio infinito vanno riconosciuti da un automa finito, devono essere formati secondo regole rigide.

Le regole furono enunciate da Kleene nel 1956; esse definiscono una classe di linguaggi detti linguaggi regolari o insiemi regolari. Kleene dimostrò che un automa finito può riconoscere un linguaggio solo se è regolare e, inoltre, che ogni linguaggio regolare può essere riconosciuto da qualche automa finito. Cosa si intenda per regolare può essere indicato brevemente (anche se in modo non rigoroso) da due regole. Primo, qualsiasi linguaggio finito è regolare e può quindi essere riconosciuto da un automa finito; dopo tutto, si potrebbe costruire una macchina con uno stato per ogni possibile espressione del linguaggio. Secondo, se un linguaggio è infinito deve essere possibile analizzare sintatticamente tutti i suoi enunciati leggendo un simbolo alla volta da sinistra a destra, ossia dall'inizio alla fine, senza mai tornare indietro o guardare in avanti. Se l'accettabilità di un simbolo dipende dalla presenza di un altro simbolo, deve trattarsi del simbolo immediatamente a sinistra.

La seconda regola è una diretta conseguenza delle limitazioni di un automa finito, che non può né prevedere i suoi stati futuri né conservare un ricordo di quelli passati; deve scegliere una transizione di stato che si basi solo sullo stato attuale e sull'attuale simbolo di input. È per questa ragione che un automa finito non può maneggiare una notazione sottrattiva per i numerali romani. Se l'espressione XI

La gerarchia di Chomsky di macchine finite e infinite

viene letta e la macchina la interpreta come 11, non può tornare a rivedere il valore quando il carattere successivo risulta essere V. Molte altre funzioni sono governate dalla stessa limitazione. Per esempio, non è possibile costruire un automa finito che legga una sequenza di cifre binarie e stabilisca se il numero di 1 è uguale al numero di 0. Analogamente, sebbene un automa finito possa sommare numeri binari, non può moltiplicarli; lascio al lettore capire perché.

Al di là degli automi finiti e dei linguaggi regolari si estende una gerarchia di macchine più potenti e di linguaggi più generali. Tale gerarchia è detta gerarchia di Chomsky, dal nome del linguista Noam Chomsky che studiò i vari linguaggi formali come possibili modelli per il linguaggio naturale. Linguaggi più generali si creano allentando le limitazioni sulle regole grammaticali degli insiemi regolari; le macchine sono costruite aggiungendo elementi di memoria al modello base a stati finiti.

La prima macchina della serie è detta macchina finita con memoria di tipo *pushdown*. Consiste di un automa finito con l'aggiunta di una matrice di memoria con una capacità infinita ma una organizzazione particolare. La memoria prende la forma di una catasta, come una catasta di vassoi in un self-service. Un elemento di informazione può essere immagazzinato solo ponendolo in cima alla catasta e per recuperarlo bisogna prima rimuovere tutti gli elementi che gli stanno sopra. In questo modo, l'ultimo elemento entrato è il primo a uscire.

Il linguaggio riconosciuto da una macchina finita con memoria di tipo *pushdown* è detto linguaggio libero da contesto. Analizzando sintatticamente i suoi enunciati, l'accettabilità di un simbolo può dipendere sia dal simbolo immediatamente a sinistra sia da quello immediatamente a destra. Questa dipendenza bidirezionale è ammissibile perché qualsiasi simbolo la cui interpretazione non possa essere decisa immediatamente può essere immagazzinato nella catasta finché l'ambiguità è risolta. Una macchina finita con memoria di tipo *pushdown*, quindi, può lavorare con numeri romani sottrattivi e può identificare espressioni con numeri uguali di 1 e 0 (o altri simboli, quali parentesi aperta e chiusa). Invece, non può individuare enunciati con numeri uguali di tre simboli (per esempio 0, 1, 2). La maggior parte dei linguaggi di programmazione sono liberi da contesto e l'analizzatore sintattico di un compilatore è di solito una macchina finita con memoria di tipo *pushdown*. Molti calcolatori hanno dispositivi hardware per organizzare parte della capacità di memoria come catasta di tipo *pushdown*. Un linguaggio di programmazione, il Forth, fa di una catasta la struttura primaria di memoria. Naturalmente, in una macchina reale una catasta non può avere lunghezza infinita.

I linguaggi liberi da contesto hanno un nome adeguato in quanto l'analisi sintattica di un simbolo può essere influenzata direttamente solo dai due simboli immediatamente adiacenti, e non dal più ampio contesto in cui esso si trova. Togliendo questa limitazione si ottiene un linguaggio sensibile al contesto e aumenta ulteriormente la difficoltà di interpretazione. Ora possono interagire simboli molto distanti uno dall'altro; nel caso peggiore non è possibile interpretare il primo simbolo di un'espressione finché non è stato letto l'ultimo. A compenso della maggiore complessità, si guadagna qualcosa in capacità d'azione. Una macchina basata su un linguaggio sensibile al contesto può stabilire se in un'espressione si trovano numeri uguali di tre simboli.

La macchina che può riconoscere un linguaggio sensibile al contesto è un automa limitato linearmente. Oltre al consueto apparato di automa finito, ha una memoria organizzata in modo che in qualsiasi momento si possa raggiungere qualsiasi locazione di immagazzinamento; è una macchina ad accesso casuale. La memoria ha una capacità finita, ma si presuppone che sia abbastanza grande da contenere qualsiasi input la macchina riceva. L'automa limitato linearmente sembra una buona approssimazione al modello di von Neumann di calcolatore digitale. Stranamente, però, i corrispondenti linguaggi di programmazione sensibili al contesto sembrano rari; evidentemente, la più semplice struttura libera da contesto ha quasi sempre sufficiente potenza espressiva.

Tutti i linguaggi descritti sopra hanno una proprietà in comune: sono detti ricorsivi. Con questa designazione si intende che si può immaginare una procedura per generare tutte le possibili «espressioni» del linguaggio in ordine di lunghezza crescente. Ne consegue che esiste un metodo per decidere se un dato enunciato di lunghezza finita è un membro del linguaggio: basta generare tutti gli enunciati fino a quella lunghezza e confrontarli.

Vi sono linguaggi che non possono soddisfare nemmeno questo standard minimo di trattabilità. C'è solo una macchina che possa riconoscerli: è l'ultima spiaggia del calcolatore, la macchina di Turing, un automa finito che può spaziare liberamente in una memoria senza limiti. Nella descrizione data da Turing, la memoria è un nastro, infinito in entrambe le direzioni e diviso in celle su cui l'apparato a stati finiti può scrivere, leggere o cancellare.

Guardando dall'elevato punto di vista della macchina di Turing, si chiariscono le relazioni tra i più modesti congegni di calcolo. L'automa limitato linearmente è semplicemente una macchina di Turing con un nastro finito. La macchina finita con memoria di tipo *pushdown* ha un nastro infinito in una direzione, ma la «testina» per leggere e scrivere sul nastro rimane sempre fissa sull'ultima cella non vuota. L'automa finito è una macchina di Turing del tutto priva di nastro.

Forse i lettori sempre a caccia di novità, ansiosi di analizzare sintatticamente linguaggi non ricorsivi, sono già usciti per acquistare una macchina di Turing. Andrebbero avvertiti che anche il calcolatore «estremo» ha le sue debolezze. Ci sono linguaggi con grammatiche così strampalate che nemmeno con una macchina di Turing si potrebbero riconoscere i loro enunciati in un tempo finito. Finora questi linguaggi hanno trovato scarso uso nel mondo delle macchine da calcolo, ma la gente riesce in qualche modo a parlarli.